UNIVERSIDADE DA CORUÑA

Genetic analyses in various razor shell species of family Pharidae, with a focus on Atlantic *Ensis*

PhD thesis / Tesis de doctorado

Departamento de Biología Celular y Molecular

Joaquín Vierna Fernández (2014)

UNIVERSIDADE DA CORUÑA

DÑA. ANA MARÍA GONZÁLEZ TIZÓN, DOCTORA EN BIOLOGÍA Y PROFESORA CONTRATADA DOCTORA EN EL ÁREA DE GENÉTICA DEL DEPARTAMENTO DE BIOLOGÍA CELULAR Y MOLECULAR DE LA UNIVERSIDADE DA CORUÑA, Y D. ANDRÉS MARTÍNEZ LAGE, DOCTOR EN BIOLOGÍA Y PROFESOR TITULAR DE UNIVERSIAD EN EL ÁREA DE GENÉTICA DEL DEPARTAMENTO DE BIOLOGÍA CELULAR Y MOLECULAR DE LA UNIVERSIDADE DA CORUÑA,

INFORMAN

QUE EL TRABAJO TITULADO "GENETIC ANALYSES IN VARIOUS RAZOR SHELL SPECIES OF FAMILY PHARIDAE, WITH A FOCUS ON ATLANTIC *ENSIS*"**, PRESENTADO POR D. JOAQUÍN VIERNA FERNÁNDEZ PARA OPTAR AL TÍTULO DE DOCTOR EN BIOLOGÍA POR LA UNIVERSIDADE DA CORUÑA, HA SIDO REALIZADO BAJO NUESTRA DIRECCIÓN. CONSIDERÁNDOLO FINALIZADO, AUTORIZAMOS SU PRESENTACIÓN Y DEFENSA.**

A CORUÑA, A 17 DE ENERO DE 2014

Fdo. Dra. Ana María González Tizón

Fdo. Dr. Andrés Martínez Lage

ACKNOWLEDGEMENTS

I would like to thank my PhD supervisors, Ana M. González Tizón and Andrés Martínez Lage for introducing me to Genetics and for their great support. Many thanks to my colleagues at the Evolutionary Biology Group, Universidade da Coruña whom I am so glad to have worked with: Alejandra Perina, David Seoane, Elvira Sahuquillo, Horacio Naveira, Manuel Pimentel, Marcelino Fuentes, Marta Vila, Miguel Vizoso, Neus Marí, Nuria Remón, Pedro Galán, Rosa García-Junco, Verónica Rojo, and Zeltia Torrecilla. I thank the people who kindly hosted me in their labs during my internships abroad: Rudo von Cosel at the Muséum national d'Histoire naturelle (Paris); K. Thomas Jensen, Jane Frydenberg, and Camilla Håkansson at the Aarhus Universitet; Emilie Egea at the Centre d'Océanologie de Marseille; and Manja Marz, Stefanie Wehner, and Marcus Lechner at the Philipps-Universität Marburg. I would like to thank their hospitality, support and interest in my thesis work. Many thanks to Ángeles Cid for her support. Thanks to Fernanda Rodríguez and Raquel Lorenzo, from the Molecular Biology Unit of the University Research Support, Services, for sequencing most of the samples generated in this thesis. Thanks to Alejandro Martínez, Antón Vizcaíno, Belén Carro, Diego Fontaneto, Joël Cuperus, Jukka Corander, Julio Parapar, Marta Pola, Naiara Albaina, Nicolas Puillandre, Paul Dansey, Rüdiger Schmelz, Soraya Rumbo, my biology friends, and my colleagues from Precarios-Galicia, for our fruitful discussions and great support. I would also like to thank my students from the various practical courses I had the chance to be in charge of at University. Thanks to the funding institutions, especially the Universidade da Coruña, the Xunta de Galicia, the European Social Fund, and the Galician Network for Conservation of Biological Diversity, to the editors and reviewers of the articles that comprise this thesis, the external reviewers and the thesis committee. Finally, I would like to thank everyone who collected samples, provided photos of razor shell specimens, and allowed the deposit of razor shells in natural history museums.

THANK YOU SO MUCH!

Joaquín Vierna, November 2013.

AGRADECIMIENTOS

A mi directora y a mi director de tesis, Ana M. González Tizón y Andrés Martínez Lage, por introducirme en el mundo de la genética y por su gran apoyo y ayuda. También a mis compañeros/as del Grupo de Investigación en Biología Evolutiva de la Universidade da Coruña (Alejandra Perina, David Seoane, Elvira Sahuquillo, Horacio Naveira, Manuel Pimentel, Marcelino Fuentes, Marta Vila, Miguel Vizoso, Neus Marí, Nuria Remón, Pedro Galán, Rosa García-Junco, Verónica Rojo y Zeltia Torrecilla), con quienes estoy encantado de haber trabajado. A las personas que me acogieron en sus laboratorios durante las estancias realizadas en mi tesis: en el Muséum national d'Histoire naturelle (París), Rudo von Cosel; en la Aarhus Universitet, K. Thomas Jensen, Jane Frydenberg y Camilla Håkansson; en el Centre d'Océanologie de Marseille, Emilie Egea; y en la Philipps-Universität Marburg, Manja Marz, Stefanie Wehner y Marcus Lechner. A todos ellos quiero agradecer su hospitalidad, apoyo e interés por este trabajo. Muchas gracias también a Ángeles Cid por su apoyo. Gracias a Fernanda Rodríguez y a Raquel Lorenzo, de la Unidad de Biología Molecular de los Servicios de Apoyo a la Investigación de la UDC, que secuenciaron la mayor parte de las muestras. A Alejandro Martínez, Antón Vizcaíno, Belén Carro, Diego Fontaneto, Joël Cuperus, Jukka Corander, Julio Parapar, Marta Pola, Naiara Albaina, Nicolas Puillandre, Paul Dansey, Rüdiger Schmelz, Soraya Rumbo, mis amigos/as de biología y mis compañeros/as de Precarios-Galicia, con los que he debatido y consultado muchas dudas, y que me han ayudado, animado y apoyado. También les doy las gracias a los/las estudiantes de Genética, Genética Humana y Métodos y Técnicas de Estudio en Genética, con los que disfruté mucho durante las clases. Agradezco también a los organismos que han financiado estas investigaciones, principalmente la Universidade da Coruña, la Xunta de Galicia, el Fondo Social Europeo y la Red Gallega de Conservación de la Diversidad Biológica, y a los editores y revisores que permitieron la publicación de los artículos científicos que conforman esta tesis, así como a los revisores externos y al tribunal de evaluación de la misma. Por último, quiero agradecer a las personas que han recolectado y enviado muestras de navajas, han hecho fotos a especímenes de colecciones malacológicas, o han permitido el depósito en museos de historia natural, de los especímenes que han sido utilizados en esta tesis.

¡MUCHÍSIMAS GRACIAS!

Joaquín Vierna, noviembre de 2013.

1 SUMMARIES

1. SUMMARIES

1.1 English

In this thesis, a genetic study was carried out in a group of marine bivalves of the family Pharidae Adams and Adams, 1858, with a focus on genus *Ensis* Schumacher, 1817. Such study has been developed within the areas of evolutionary genetics, population genetics, cytogenetics, species delimitation, and DNA barcoding.

The razor shell species under study were selected due to the commercial importance that many *Ensis* species have in various European and American regions, including Galicia. Moreover, their amphi-atlantic distribution and their occurrence in some areas of the Pacific coast of the USA, Peru, and Chile makes them interesting target species for phylogeographic studies. Finally, these species were selected due to the existing problems on their morphological identifications, in order to develop molecular and cytogenetic tools that complement morphometric studies.

The methodology varies along each of the chapters of the thesis, but can be summarised as follows: in CHAPTER 4.1.1, 5S ribosomal DNA and U1 small nuclear DNA, both multicopy families, were studied following a PCR amplification, cloning, and sequencing procedure. All the sequences obtained in the wet lab were subsequently analysed by several bioinformatic tools. In CHAPTER 4.1.2, in which 97 metazoan species were studied, the methodology employed was different, because genome project databases were used, together with high throughput bioinformatic tools specifically designed for this study. In CHAPTER 4.2 a cytogenetic study was carried out in which fluorescence and light microscopy were used, along with hybridisation of labelled fluorescent probes on razor shell chromosomes, that served as chromosomal markers. Karyotypes were obtained by measuring the chromosomes and by sorting them out by size and centromeric index. Finally, in CHAPTERS 4.3 and 4.4 various nuclear and mitochondrial genome regions were used as molecular markers, being amplified and sequenced (and in some cases, cloned prior to sequencing). These molecular markers are fragments of mitochondrial genes cytochrome oxidase subunit I and 16S ribosomal DNA; fragments of nuclear genes adenine nucleotide translocase and 18S ribosomal DNA; and the nuclear region encompassing the ribosomal spacers ITS1 and ITS2, and the 5.8S ribosomal DNA gene. Some of these markers were employed at the population level, at the species level, or at both the population and species levels.

In CHAPTER 4.1.1 the linkage between 5S ribosomal DNA and U1 small nuclear DNA was studied. Linked units between repeats of these multigene families were characterised in 10 Pharidae species (from four different genera). We obtained clones containing partial or complete repeats of both

genes in which the RNA coding regions displayed the same orientation. A comprehensive collection of 5S ribosomal DNA clones of various razor shells both linked and non-linked to U1 small nuclear DNA was obtained. The secondary structure of RNA coding regions was predicted, and the conserved upstream and downstream elements were characterised. The analysis of the nontranscribed spacers (NTSs) of the 5S ribosomal DNA showed that some of them were more closely related to NTS from other species than to NTS from the species they were retrieved from, suggesting ancestral polymorphism and long-term evolution by the birth-and-death process. Nucleotide conservation within the functional regions suggested that purifying selection, in addition to unequal crossing-overs and gene conversions, were involved in the evolution of these multigene families. Taking this and other studies into account, the possible mechanisms by which both multigene families could have become linked in the Pharidae were discussed. The reason why 5S ribosomal DNA is often found linked to other multigene families seems to be due to stochastic processes within genomes in which its high copy number is determinant.

In CHAPTER 4.1.2 the former study was broadened up to 97 metazoan species, whose genomes were sequenced and available in the international databases. 5S ribosomal DNA sequences from the main metazoan clades were systematically analysed and compared, using bioinformatic methods. After having performed a filtering of candidate sequences, 12766 putatively functional copies were obtained, which enabled us to identify some general features of 5S ribosomal DNA in animals. We also showed that each mammal species analysed has a highly conserved copy of 5S ribosomal DNA (that we named *housekeeping*) as well as many other variable copies. The NTS sequences, the genomic organisation of 5S ribosomal DNA, and its evolution, were analysed in detail. Our results confirmed the existence of paralogous copies in 58 genomes. Moreover, a flexible genomic organisation of this multigene family was found in metazoans. The existence of heterogeneous clusters of 5S ribosomal DNA (composed of similar coding regions and divergent NTSs) in many species supports the hypothesis of an exchange of 5S ribosomal DNA from one locus to another. We also performed a detailed analysis of the degree of conservation of internal and upstream promoter regions, as well as downstream termination signals. Finally, we described an statistical method to analyse the linkage among non-coding RNA multigene families, that failed to detect any stable linkage between 5S ribosomal DNA and other multigene families throughout the evolution of metazoans.

In the following chapter (4.2) a cytogenetic study of the European razor shell *Ensis minor* (Chenu, 1843) and of the American *E. directus* (Conrad, 1843) was carried out. The diploid chromosome number of both species was 38, and remarkable differences were unveiled in their

karyotypes. *E. minor* has four pairs of metacentric chromosomes, one metacentric-submetacentric, five submetacentrics, one subtelocentric, and eight telocentrics. In turn, *E. directus* has three pairs of metacentric chromosomes, two metacentric-submetacentrics, six submetacentrics, six subtelocentrics, and two telocentrics. Fluorence *in situ* hybridisation using a major ribosomal genes probe localised these genes on a submetacentric chromosome pair in both species. Hybridisation using a 5S ribosomal DNA probe yielded a weak chromosomal signal in *E. minor* and no signal at all in *E. directus*, supporting a more dispersed organisation of the 5S ribosomal DNA, compared to the major ribosomal genes, in these species. The vertebrate telomeric probe $(TTAGGG)_n$ hibridised to both telomers of each chromosome, without yielding any interstitial signal. In this study, a comparative karyological analysis was carried out among the four *Ensis* studied so far, the European species *E. minor*, *E. siliqua* (Linné, 1758), and *E. magnus* Schumacher, 1817, and the American species *E. directus*. European *Ensis* showed more similarities between them than between either of them and *E. directus*. Moreover, clear karyotypic differences were found between the morphologically similar species *E. minor* and *E. siliqua* in the number of telocentric and subtelocentric chromosome pairs.

In CHAPTER 4.3 we obtained the levels of current genetic variation of populations of the razor shell *E. directus* (Mollusca: Bivalvia: Pharidae) at both the native (North America) and introduced (Europe) distributional ranges by using nuclear and mitochondrial sequences. We expected less genetic variation in the introduced range, specially considering the frequent mass mortality events observed in Europe since the introduction of the species in 1978. Nevertheless, we found higher variation in Europe. Results were discussed in the light of the possible influence of temporal fluctuations of genetic variation, the limited effect of random genetic drift, and the possibility of multiple introductions. Interestingly, the hypothesis of multiple introductions contrasts with the gradual colonisation of the European coastline by *E. directus*, but it is supported by the intensity of traffic flow in the Atlantic. Finally, genetic and morphometric evidences strongly supported that individuals from a population from Newfoundland (Canada) belonged to a new species. This new *Ensis* was formally described and named *E. terranovensis* n.sp.

In CHAPTER 4.4 a species delimitation and DNA barcoding study was performed, in which Atlantic *Ensis* species were considered with the aim of testing whether extant Atlantic morphospecies were different evolutionary lineages. In this work, we studied 109 specimens belonging to the nine *Ensis* species known to occur in the Atlantic, and surveyed genetic variation at four nuclear and two mitochondrial genomic regions.

The phylogenetic analyses carried out supported the reciprocal monophyly of *Ensis* species at each

side of the Atlantic. In Europe four separate lineages were found, corresponding to *E. magnus* Schumacher, 1817, *E. ensis* (Linné, 1758), *E. minor*, and *E. siliqua*. The co-occurrence of *E. minor* and *E. siliqua* along the NW Iberian coast was also demonstrated. A quite high degree of divergence was unveiled among individuals of *E. macha* (Molina, 1792) from Chilean and Argentinean populations, suggesting incipient speciation. Moreover, the occurrence of *E. directus* off the northern coast of Florida was confirmed. Among the genomic regions analysed, the fragment of the cytochrome oxidase subunit I is proposed as the more suitable DNA barcode for the identification of Atlantic *Ensis* species.

The most relevant contributions of this thesis are:

- We demonstrated that 5S ribosomal DNA copies are linked to U1 small nuclear DNA copies in at least 10 Pharidae species, from four different genera. Both multigene families were characterised at the nucleotide level, and the first Bivalvia U1 small nuclear DNA sequences were reported.
- The 5S ribosomal DNA diversity was characterised at a broad evolutionary scale, that is, in 97 metazoan species.
- The razor shells *E. minor* y *E. directus* were characterised at the cytogenetic level, and some interesting insights regarding the taxonomy of *Ensis* species and the organisation of 5S ribosomal DNA and the major ribosomal genes were obtained.
- Current genetic variation of *E. directus* populations in native and introduced ranges was reported, and a new species from Newfoundland (Canada) was discovered, described, and named *E. terranovensis*.
- The taxonomic status of the extant Atlantic *Ensis* was clarified, and the COI region was proposed as the most suitable for the identification of these species by DNA barcoding.

1.2 Spanish

En esta tesis se ha realizado un estudio genético de un grupo de bivalvos marinos de la familia Pharidae Adams y Adams, 1858, y principalmente de las navajas del género *Ensis* Schumacher, 1817. El estudio se desarrolla dentro de las áreas de la genética evolutiva, genética poblacional, citogenética, delimitación de especies y *DNA barcoding*.

Las especies de navaja utilizadas fueron seleccionadas debido al interés comercial que muchas

especies de *Ensis* tienen en varias regiones de Europa y América, incluyéndose también el litoral de Galicia. Además, su distribución a ambos lados del Atlántico y en zonas de la costa pacífica de los Estados Unidos, El Perú y Chile hace que sean de interés en estudios filogeográficos. Por último, los problemas en la identificación de las especies a partir de caracteres morfológicos hicieron de este grupo de organismos un buen candidato para el desarrollo de herramientas moleculares y citogenéticas que complementasen los estudios de morfometría.

La metodología utilizada varía en cierto modo a lo largo de cada uno de los capítulos que conforman los resultados de la tesis, aunque puede resumirse como sigue: en el CAPÍTULO 4.1.1, el ADN ribosomal 5S y el ADN pequeño nuclear U1, que conforman familias génicas de copia múltiple, fueron estudiados mediante un procedimiento de amplificación por PCR, clonación y secuenciación. Todas estas secuencias obtenidas experimentalmente en el laboratorio fueron posteriormente analizadas mediante numerosas herramientas bioinformáticas. En el CAPÍTULO 4.1.2, en el que se estudiaron 97 especies de metazoos, la metodología utilizada fue diferente, ya que se utilizaron bases de datos de proyectos genoma y herramientas bioinformáticas de alta capacidad, algunas de las cuales fueron diseñadas expresamente para este trabajo. En el CAPÍTULO 4.2 se realizó un estudio citogenético en el que se utilizaron técnicas de microscopía en campo claro y fluorescente, así como la hibridación de sondas fluorescentes sobre los cromosomas de las navajas, que fueron utilizadas como marcadores cromosómicos. Se realizó también la medición y ordenación de los cromosomas por tamaño e índice centromérico, para elaborar los cariotipos. Por último, en los CAPÍTULOS 4.3 y 4.4 se utilizaron como marcadores moleculares diferentes regiones de los genomas nuclear y mitocondrial, que han sido amplificadas mediante PCR, siendo a veces clonadas, previamente a su secuenciación. Estos marcadores moleculares son fragmentos de los genes mitocondriales citocromo oxidasa subunidad I y ADN ribosomal 16S; fragmentos de los genes nucleares adenina nucleótido translocasa y ADN ribosomal 18S; y la región nuclear que contiene los espaciadores ribosomales ITS1 e ITS2, y el gen ribosomal 5.8S. Algunos de ellos se han empleado solo a nivel poblacional, algunos solo a nivel de especie, mientras que otros se han empleado a ambos niveles.

En el CAPÍTULO 4.1.1 se estudió el ligamiento entre el ADN ribosomal 5S y el ADN nuclear pequeño U1. Se describieron unidades de ligamiento entre ambas familias multigénicas en 10 especies (de cuatro géneros diferentes) de la familia Pharidae. Se obtuvo un número de clones que contenían repeticiones parciales o completas de ambos genes, en las que las regiones codificantes mostraron la misma orientación. Se obtuvo una completa colección de clones de ADN ribosomal 5S de varias especies de navajas, tanto ligados como no ligados al ADN nuclear pequeño U1. También

se predijo la estructura secundaria de las regiones codificantes, y se caracterizó los elementos conservados de las regiones aguas arriba y aguas abajo de las mismas. El análisis de los espaciadores no transcritos (NTS) del ADN ribosomal 5S mostró que algunos de ellos estaban evolutivamente más relacionados con NTS de otras especies que con aquellos de la misma, sugiriendo polimorfismo ancestral y evolución a largo plazo mediante el proceso de *birth-and-death*. La conservación nucleotídica dentro de las regiones funcionales sugirió que la selección purificadora, además de los entrecruzamientos desiguales y las conversiones génicas, estuvieron implicados en la evolución de estas familias multigénicas. Considerando este estudio y otros previos, se discutieron los posibles mecanismos mediante los cuales ambas familias multigénicas han podido establecerse en unidades de ligamiento en el linaje de los Pharidae. La razón por la que el ADN ribosomal 5S se encuentra a menudo ligado a otras familias multigénicas parece ser el resultado de procesos estocásticos dentro de los genomas en el que su alto número de copia sería determinante.

En el CAPÍTULO 4.1.2 se amplió el estudio anterior a 97 especies de metazoos, cuyos genomas estaban secuenciados y disponibles en bases de datos internacionales. Se analizaron y compararon sistemáticamente secuencias de ADN ribosomal 5S en especies de los principales clados de metazoos, usando métodos bioinformáticos. Tras haber realizado un filtrado de las secuencias candidatas, se obtuvieron 12766 copias putativamente funcionales que nos permitieron identificar algunas características generales del ADN ribosomal 5S de los animales. También se muestra que cada especie de mamífero analizada tiene una copia altamente conservada de ADN ribosomal 5S (que denominamos *housekeeping*) así como muchas otras copias más variables. Se analizó en detalle los NTS, la organización de esta familia multigénica en el genoma y su evolución. Nuestros resultados confirmaron la existencia de copias parálogas en 58 genomas. También fue evidente una organización flexible dentro del genoma de los animales. La existencia de agrupaciones heterogéneas de copias de ADN ribosomal 5S (compuestas de regiones codificantes similares y de NTS divergentes) en muchas especies apoya la hipótesis de un intercambio de ADN ribosomal 5S de un locus a otro del genoma. Además, obtuvimos un análisis detallado del grado de conservación evolutiva de las regiones promotoras internas, aguas arriba y aguas abajo en animales. Por último, describimos un método estadístico para analizar el ligamiento entre familias multigénicas codificadoras de ARN, que sin embargo no obtuvo ningún resultado de ligamiento estable entre el ADN ribosomal 5S y otras familias a lo largo de la evolución de los metazoos.

En el siguiente capítulo (4.2) se realizó un estudio citogenético de la navaja europea *Ensis minor* (Chenu, 1843) y de la americana *E. directus* (Conrad, 1843). Se vio que ambas tienen un número

cromosómico diploide de 38 y notables diferencias cariotípicas. *E. minor* tiene cuatro pares de cromosomas metacéntricos, uno metacéntrico-submetacéntrico, cinco submetacéntricos, uno subtelocéntrico y ocho telocéntricos. En cambio *E. directus* tiene tres pares metacéntricos, dos metacéntricos-submetacéntricos, seis submetacéntricos, seis subtelocéntricos y dos telocéntricos. La hibridación *in situ* fluorescente usando una sonda de genes ribosomales mayores localizó estos genes en un par submetacéntrico en ambas especies. La hibridación con ADN ribosomal 5S produjo una señal cromosómica débil en *E. minor* y ninguna en *E. directus*, apoyando una organización más dispersa de esta familia multigénica, comparada con la de los genes ribosomales mayores. La sonda telomérica de vertebrados $(TTAGGG)_n$ hibridó en ambos telómeros de cada cromosoma, sin señales intersticiales. Además, en este trabajo se realizó un estudio cariológico comparado de las cuatro *Ensis* analizadas hasta la fecha, las europeas *E. minor*, *E. siliqua* (Linné, 1758) y *E. magnus* Schumacher, 1817, y la americana *E. directus*. Las especies europeas mostraron más similitudes entre ellas que con *E. directus*. Además, se encontraron diferencias cariotípicas claras entre las especies morfológicamente similares *E. minor* y *E. siliqua*, en el número de pares cromosómicos telocéntricos y subtelocéntricos.

En el CAPÍTULO 4.3 obtuvimos los niveles de variación genética actual de poblaciones de la navaja *E. directus* (Mollusca: Bivalvia: Pharidae) en los rangos de distribución nativo (América del Norte) e introducido (Europa) usando secuencias nucleares y mitocondriales. Esperábamos menor variación en el rango de distribución introducido, sobre todo considerando los frecuentes episodios de mortalidad en masa observados en Europa desde la introducción de la especie en 1978. Sin embargo, encontramos mayor variación en Europa. En este trabajo los resultados se comentaron a la luz de la posible influencia de incrementos o reducciones temporales de la variación genética, del efecto limitado de la deriva genética aleatoria y de posibles introducciones múltiples. Curiosamente, la hipótesis de las introducciones múltiples contrasta con la colonización gradual de la costa europea por parte de *E. directus*, pero es apoyada por la intensidad del tráfico transoceánico en el Atlántico. Por último, evidencias genéticas y morfométricas apoyaron claramente que los individuos de una población analizada de Terranova (Canadá) pertenecían a una especie nueva, desconocida hasta la fecha. Esta nueva *Ensis* se describió formalmente en este capítulo y fue denominada *E. terranovensis* n.sp.

En el último capítulo (4.4) se realizó un trabajo de delimitación de especies y *DNA barcoding* en las *Ensis* del Atlántico, en el que se estudió si las morfoespecies actualmente descritas eran linajes evolutivos diferentes. En este trabajo estudiamos 109 especímenes pertenecientes a nueve especies de *Ensis* (todas las especies actuales del Atlántico) y en ellas analizamos la variación nucleotídica

en cuatro regiones nucleares y en dos regiones mitocondriales. Los análisis filogenéticos realizados apoyan la monofilia recíproca de estas especies a cada lado del océano Atlántico. En Europa se encontraron cuatro linajes claramente diferenciados, que se correspondieron con las especies *E. magnus*, *E. ensis* (Linné, 1758), *E. minor* y *E. siliqua*, demostrándose, además, que *E. minor* y *E. siliqua* conviven en la costa NO de la Península Ibérica. Un grado de divergencia bastante relevante se apreció entre individuos de *E. macha* (Molina, 1792) muestreados en Chile y en Argentina, lo cual sugiere especiación incipiente. Además, se confirmó la presencia de *E. directus* al norte de Florida. De entre las regiones genómicas analizadas, se sugiere el fragmento de la citocromo oxidasa subunidad I para ser utilizada en identificación mediante *DNA barcoding*.

Las contribuciones más relevantes de esta tesis son las siguientes:

- Se demostró que existen copias de ADN ribosomal 5S ligadas a copias de ADN pequeño nuclear U1 en los genomas de al menos 10 especies de la familia Pharidae, de cuatro géneros diferentes. Además de caracterizar, a nivel nucleotídico, ambas familias multigénicas, las secuencias obtenidas del ADN pequeño nuclear U1 son las primeras en la Clase Bivalvia.
- Se caracterizó por vez primera la diversidad del ADN ribosomal 5S en una escala evolutiva amplia, es decir, en 97 especies de metazoos.
- Se caracterizaron las especies *E. minor* y *E. directus* a nivel citogenético, obteniéndose conclusiones aplicadas a la taxonomía de estas especies, y a la organización genómica del ADN ribosomal 5S y de los genes ribosomales mayores.
- Se estudió la variación genética de poblaciones de *E. directus* en los rangos de distribución nativo e introducido, y se descrubrió y describió una nueva especie en Terranova (Canadá), a la que se le llamó *E. terranovensis*.
- Se clarificó el estatus taxonómico de las especies actuales de *Ensis* en el Atlántico, y se definió que la región COI es adecuada para la identificación de estas especies mediante *DNA barcoding*.

1.3 Galician

Nesta tese realizouse un estudo xenético dun grupo de bivalvos mariños da familia Pharidae Adams e Adams, 1858, e principalmente das navallas do xénero *Ensis* Schumacher, 1817. O estudo desenvólvese dentro das áreas da xenética evolutiva, xenética poboacional, citoxenética,

delimitación de especies e *DNA barcoding*.

As especies de navalla utilizadas foron seleccionadas debido á interese comercial que moitas especies de *Ensis* teñen en varias rexións de Europa e América, incluíndose tamén o litoral de Galicia. Ademais, a súa distribución a ambos os dous lados do Atlántico e en zonas da costa pacífica dos Estados Unidos, O Perú e Chile fai que sexan de interese en estudos filoxeográficos. Para rematar, os problemas na identificación das especies a partir de caracteres morfolóxicos fixeron deste grupo de organismos un bo candidato para o desenvolvemento de ferramentas moleculares e citoxenéticas que complementasen os estudos de morfometría.

A metodoloxía utilizada varía en certo xeito ao longo de cada un dos capítulos que conforman os resultados da tese, aínda que pode resumirse como segue: no CAPÍTULO 4.1.1, o ADN ribosomal 5S e o ADN pequeno nuclear U1, que conforman familias xénicas de copia múltiple, foron estudados mediante un procedemento de amplificación por PCR, clonaxe e secuenciación. Todas estas secuencias obtidas experimentalmente no laboratorio foron posteriormente analizadas mediante numerosas ferramentas bioinformáticas. No CAPÍTULO 4.1.2, no que se estudaron 97 especies de metazoos, a metodoloxía utilizada foi diferente, xa que se utilizaron bases de datos de proxectos xenoma e ferramentas bioinformáticas de alta capacidade, algunhas das cales foron deseñadas expresamente para este traballo. No CAPÍTULO 4.2 realizouse un estudo citoxenético no que se utilizaron técnicas de microscopía en campo claro e fluorescente, así como a hibridación de sondas fluorescentes sobre os cromosomas das navallas, que foron utilizadas como marcadores cromosómicos. Realizouse tamén a medición e ordenación dos cromosomas por tamaño e índice centromérico, para elaborar os cariotipos. Para rematar, nos CAPÍTULOS 4.3 e 4.4 utilizáronse como marcadores moleculares diferentes rexións dos xenomas nuclear e mitocondrial, que foron amplificadas mediante PCR, sendo ás veces clonadas, previamente á súa secuenciación. Estes marcadores moleculares son fragmentos dos xenes mitocondriais citocromo oxidasa subunidade I e ADN ribosomal 16S; fragmentos dos xenes nucleares adenina nucleótido translocasa e ADN ribosomal 18S; e a rexión nuclear que contén os espaciadores ribosomais ITS1 e ITS2, e o xene ribosomal 5.8S. Algúns deles empregáronse só a nivel poboacional, algúns só a nivel de especie, mentres que outros se empregaron a ambos niveis.

No CAPÍTULO 4.1.1 estudouse o ligamento entre o ADN ribosomal 5S e o ADN nuclear pequeno U1. Describíronse unidades de ligamento entre ambas familias multixénicas en 10 especies (de catro xéneros diferentes) da familia Pharidae. Obtívose un número de clons que contiñan repeticións parciais ou completas de ambos xenes, nas que as rexións codificantes mostraron a mesma orientación. Obtívose unha completa colección de clons de ADN ribosomal 5S de varias especies de

navallas, tanto ligados como non ligados ao ADN nuclear pequeno U1. Tamén se predeciu a estrutura secundaria das rexións codificantes, e caracterizáronse os elementos conservados das rexións augas arriba e augas abaixo das mesmas. A análise dos espaciadores non transcritos (NTS) do ADN ribosomal 5S mostrou que algúns deles estaban evolutivamente máis relacionados con NTS doutras especies que con aqueles da mesma, suxerindo polimorfismo ancestral e evolución a longo prazo mediante o proceso de *birth-and-death*. A conservación nucleotídica dentro das rexións funcionais suxeriu que a selección purificadora, ademais dos entrecruzamentos desiguais e as conversións xénicas, estivo implicada na evolución destas familias multixénicas. Considerando este estudo e outros previos, discutíronse os posibles mecanismos mediante os cales ambas familias multixénicas puideron establecerse en unidades de ligamento na linaxe dos Pharidae. A razón pola que o ADN ribosomal 5S atópase a miúdo ligado a outras familias multixénicas parece ser o resultado de procesos estocásticos dentro dos xenomas no que o seu alto número de copia sería determinante.

No CAPÍTULO 4.1.2 ampliouse o estudo anterior a 97 especies de metazoos, cuxos xenomas estaban secuenciados e dispoñibles en bases de datos internacionais. Analizáronse e compararon sistemáticamente secuencias de ADN ribosomal 5S en especies dos principais clados de metazoos, usando métodos bioinformáticos. Tras realizar un filtrado das secuencias candidatas, obtivéronse 12766 copias putativamente funcionais que nos permitiron identificar algunhas características xerais do ADN ribosomal 5S dos animais. Tamén se mostra que cada especie de mamífero analizada ten unha copia altamente conservada de ADN ribosomal 5S (que denominamos *housekeeping*) así como moitas outras copias máis variables. Analizouse en detalle os NTS, a organización desta familia multixénica no xenoma e a súa evolución. Os nosos resultados confirmaron a existencia de copias parálogas en 58 xenomas. Tamén foi evidente unha organización flexible dentro do xenoma dos animais. A existencia de agrupaciones heteroxéneas de copias de ADN ribosomal 5S (compostas de rexións codificantes similares e de NTS diverxentes) en moitas especies apoia a hipótese dun intercambio de ADN ribosomal 5S dun locus a outro do xenoma. Ademais, obtivemos unha análise detallada do grao de conservación evolutiva das rexións promotoras internas, augas arriba e augas abaixo en animais. Para rematar, describimos un método estatístico para analizar o ligamento entre familias multixénicas codificadoras de ARN, que con todo non obtivo ningún resultado de ligamento estable entre o ADN ribosomal 5S e outras familias ao longo da evolución dos metazoos.

No seguinte capítulo (4.2) realizouse un estudo citoxenético da navalla europea *Ensis minor* (Chenu, 1843) e da americana *E. directus* (Conrad, 1843). Viuse que ambas teñen un número cromosómico diploide de 38 e notables diferenzas cariotípicas. *E. minor* ten catro pares de

cromosomas metacéntricos, un metacéntrico-submetacéntrico, cinco submetacéntricos, un subtelocéntrico e oito telocéntricos. En cambio, *E. directus* ten tres pares metacéntricos, dous metacéntricos-submetacéntricos, seis submetacéntricos, seis subtelocéntricos e dous telocéntricos. A hibridación *in situ* fluorescente usando unha sonda de xenes ribosomais maiores localizou estes xenes nun par submetacéntrico en ambas especies. A hibridación con ADN ribosomal 5S produciu un sinal cromosómico débil en *E. minor* e ningún en *E. directus*, apoiando unha organización máis dispersa desta familia multixénica, comparada coa dos xenes ribosomais maiores. A sonda telomérica de vertebrados (TTAGGG)n hibridou en ambos telómeros de cada cromosoma, sen sinais intersticiais. Ademais, neste traballo realizouse un estudo cariolóxico comparado das catro *Ensis* analizadas ata a data, as europeas *E. minor*, *E. siliqua* (Linné, 1758) e *E. magnus* Schumacher, 1817, e a americana *E. directus*. As especies europeas mostraron máis similitudes entre elas que con *E. directus*. Ademais, atopáronse diferenzas cariotípicas claras entre as especies morfolóxicamente similares *E. minor* e *E. siliqua*, no número de pares cromosómicos telocéntricos e subtelocéntricos.

No CAPÍTULO 4.3 obtivemos os niveis de variación xenética actual de poboacións da navalla *E. directus* (Mollusca: Bivalvia: Pharidae) nos rangos de distribución nativo (América do Norte) e introducido (Europa) usando secuencias nucleares e mitocondriais. Esperabamos menor variación no rango de distribución introducido, sobre todo considerando os frecuentes episodios de mortalidade en masa observados en Europa desde a introdución da especie en 1978. Con todo, atopamos maior variación en Europa. Neste traballo os resultados comentáronse á luz da posible influencia de incrementos ou reducións temporais da variación xenética, do efecto limitado de deriva xenética aleatoria e de posibles introducións múltiples. Curiosamente, a hipótese das introducións múltiples contrasta coa colonización gradual da costa europea por parte de *E. directus*, pero é apoiada pola intensidade do tráfico transoceánico no Atlántico. Para rematar, evidencias xenéticas e morfométricas apoiaron claramente que os individuos dunha poboación analizada de Terranova (Canadá) pertencían a unha especie nova, descoñecida ata a data. Esta nova *Ensis* describiuse formalmente neste capítulo e foi denominada *E. terranovensis* n.sp.

No último capítulo (4.4) realizouse un traballo de delimitación de especies e *DNA barcoding* nas *Ensis* do Atlántico, no que se estudou se as morfoespecies actualmente descritas eran linaxes evolutivas diferentes. Neste traballo estudamos 109 especímenes pertencentes a nove especies de *Ensis* (todas as especies actuais do Atlántico) e nelas analizamos a variación nucleotídica en catro rexións nucleares e en dúas rexións mitocondriais. As análises filoxenéticas realizadas apoian a monofilia recíproca destas especies a cada lado do océano Atlántico. En Europa atopáronse catro

linaxes claramente diferenciadas, que se corresponderon coas especies *E. magnus, E. ensis* (Linné, 1758), *E. minor* e *E. siliqua*, demostrándose, ademais, que *E. minor* e *E. siliqua* conviven na costa NO da Península Ibérica. Un grado de diverxencia bastante relevante apreciouse entre individuos de *E. macha* (Molina, 1792) de Chile e Arxentina, o cal suxire especiación incipiente. Ademais, confirmouse a presenza de *E. directus* ao norte da Florida. De entre as rexións xenómicas analizadas, suxírese o fragmento da citocromo oxidasa subunidade I para ser utilizada en identificación mediante *DNA barcoding*.

As contribucións máis relevantes desta tese son as seguintes:

- Demostrouse que existen copias de ADN ribosonal 5S ligadas a copias de ADN pequeno nuclear U1 nos xenomas de polo menos 10 especies da familia Pharidae, de catro xéneros diferentes. Ademais de caracterizar, a nivel nucleotídico, ambas familias multixénicas, as secuencias obtidas do ADN pequeno nuclear U1 son as primeiras na Clase Bivalvia.
- Caracterizouse por vez primeira a diversidade do ADN ribosomal 5S nunha escala evolutiva ampla, é dicir, en 97 especies de metazoos.
- Caracterizáronse as especies *E. minor* e *E. directus* a nivel citoxenético, obténdose conclusións aplicadas á taxonomía destas especies, e á organización xenómica do ADN ribosomal 5S e dos xenes ribosomais maiores.
- Estudouse a variación xenética de poboacións de *E. directus* nos rangos de distribución nativo e introducido, e se descubriu e describiu unha nova especie en Terranova (Canadá), á que se lle chamou *E. terranovensis*.
- Clarificouse o estatus taxonómico das especies actuais de *Ensis* no Atlántico, e definiuse que a rexión COI é adecuada para a identificación destas especies mediante *DNA barcoding*.

2 INTRODUCTION

2. INTRODUCTION

2.1 Organisation of the thesis

In this PhD thesis we have studied a group of marine bivalve molluscs, commonly referred to as razor shells. In terms of methodology, both cytogenetic and molecular genetic approaches were followed in order to answer questions mainly related to evolutionary genetics, taxonomy, and ecology.

Since the thesis has been written as a compendium of research articles, each of them has its own introductory section. Therefore, in this general introduction, we will give an overview on the organisation of the thesis, and will provide some useful information that, due to editorial reasons, could not have been included in the articles themselves.

For the sake of clarity, the articles haven been organised according to the date they were publised, and according to their topic.

In CHAPTER 4.1.1, we characterise the linked units of two multigene families, the 5S ribosomal DNA (5S rDNA) and the U1 small nuclear DNA (U1 snDNA), in several razor shell species from four different genera (*Ensis* Schumacher, 1817, *Siliqua* Mühlfeld, 1811, *Ensiculus* H Adams, 1860, and *Pharus* Gray, 1840). We provide a comprehensive collection of razor shell 5S rDNA clones, both with linked and nonlinked organisation, and the first bivalve U1 snDNA sequences.

In CHAPTER 4.1.2, we study nucleotide diversity in the 5S rDNA region of 97 metazoan species, using genome-project data, and from an evolutionary perspective. We also study the linkage of 5S rDNA to other non-coding RNA families, in metazoans.

In CHAPTER 4.2, a cytogenetic study is carried out in two *Ensis* species, *E. directus* (Conrad, 1843) and *E. minor* (Chenu, 1843).

In CHAPTER 4.3 we focus on *E. directus* and study its populations along the native and introduced ranges, following a population genetic approach.

In CHAPTER 4.4, we use different genomic regions in order to delimit species boundaries in extant Atlantic *Ensis*, and to enable molecular identifications by DNA barcoding.

In CHAPTER 5, we report some errors observed after the publication of some of the research articles that are part of this thesis.

In CHAPTER 6, we discuss the implications of this thesis for evolutionary genetic studies, and for improving the management of *Ensis* populations.

Finally, in CHAPTER 7 we summarise the main conclusions.

The 'Appendix' section (CHAPTER 10), consists of two more articles that were not included in the main part of this thesis. The data of CHAPTER 10.1 were re-analysed in CHAPTER 4.1.1, whereas one of the main conclusions of CHAPTER 10.2 is not valid after the publication of CHAPTER 4.3 and therefore it was also excluded.

2.2 Introduction to razor shells

2.2.1 Razor shells are an economically important natural resource

'Razor shell' is the common *British English* name to refer to a group of marine bivalves classified into the superfamily Solenoidea Lamarck, 1809 (details on taxonomic nomenclature are given below). These animals are often called 'razor clams' in *American English*.

Razor shells are popular in many regions since their elongated shells are frequently found on beaches and collected by amateur malacologists, and many species are an economically important natural resource.

In some European countries, *Ensis* species are considered a delicacy. For instance, in Galicia (NW Spain), 344 tons of *E. magnus* Schumacher, 1817, and 87.6 tons of *E. siliqua* were sold in the wholesale market in 2012, representing an income of € 2.86 million for both species (Plataforma Tecnolóxica da Pesca 2013).

In the same way, in Italy, the *E. minor* fishery is one of the most important ones in the Gulf of Trieste (Del Piero and Dacaprile 1998).

In the western United States the species *Siliqua patula* (Dixon, 1789) is a popular shellfish which is often gathered by recreational harvesters. Specifically, in the estate of Alaska, about one million of those razor shells are fished each year as recreational fishing (Alaska Department of Fish and Game 2010).

The South American species *E. macha* (Molina, 1792) is one of the most important razor shells in volume of captures, even though its fisheries are exploited since quite recently (from 1988 onwards). In Chile, almost 6000 tons were landed in 1999 (Barón et al. 2004), and 1400 tons during 2005 (Ariz Abarca et al. 2007).

However, the most important razor shell according to its economic value is *Sinonovacula constricta* (Lamarck, 1818), which is cultured in China, where its production reached 345000 tons in 1996

(Guo et al. 1999). This species is one of the four most popular edible clam species in that country (Wang et al. 2010).

2.2.2 Taxonomy and systematics of *Ensis* species

The superfamily Solenoidea is an infaunal soft bottom dwelling bivalve group consisting of the two marine families, Solenidae Lamarck, 1809, and Pharidae Adams and Adams, 1858 (Cosel 1993). The Pharidae are known since the Late Cretaceous (Cosel 1990) (99.6 - 65.5 mya) and comprise at least two subfamilies, the Pharinae Adams and Adams, 1858 with genera *Pharus*, *Nasopharus* Cosel, 1993, and *Sinupharus* Cosel, 1993, and the Cultellinae Davies, 1935, with *Cultellus* Schumacher, 1817, *Sinucultellus* Cosel, 1993, *Afrophaxas* Cosel, 1993, *Phaxas* Leach in Gray, 1852, *Ensiculus*, and *Ensis.*

The monophyly of these subfamilies has not been tested using molecular data so far, and therefore we should be cautious regarding the validity of this systematic arrangement. In fact, the related genera *Siliqua*, *Pharella* Gray, 1854, *Orbicularia* Deshayes, 1850, and *Sinonovacula* Prashad, 1924 may be classified into separate subfamilies from the Cultellinae, according to morphological characters (Cosel 1993).

Table 1 Taxonomic arrangement of extant *Ensis* spp. and geographic regions where these taxa are native to.

Taxa	Synonyms	Distribution
Ensis ensis (Linné, 1758)	*Ensis phaxoides* Van Urk, 1964; *Ensis sicula* Van Urk, 1964	European coasts (Atlantic and Mediterranean); north of Africa
Ensis minor (Chenu, 1843)	*Ensis siliqua* var. *minor* Chenu, 1843	European coasts (Atlantic and Mediterranean); north of Africa
Ensis magnus Schumacher, 1817	*Ensis arcuatus* (Jeffreys, 1865); *Ensis arcuatus* var. *ensoides* Van Urk, 1964; *Ensis arcuatus* var. *norvegica* Van Urk, 1964	Atlantic European coasts
Ensis siliqua (Linné, 1758)		Atlantic European coasts
Ensis goreensis (Clessin, 1888)		Tropical west Africa
Ensis directus (Conrad, 1843)	*Ensis americanus* (Gould, 1870)	Atlantic north America
Ensis megistus megistus Pilsbry and McGinty, 1943	*Ensis minor megistus* Pilsbry and McGinty, 1943; *Ensis coseli megistus* Pilsbry and McGinty, 1943	Atlantic north America
Ensis megistus coseli Vierna, 2013	*Ensis minor* Dall, 1899; *Ensis coseli megistus* Vierna, 2013	Atlantic north America
Ensis terranovensis Vierna and Martínez-Lage, 2012		Atlantic north America
Ensis macha (Molina, 1792)	*Ensis luzonicus* Dunker, 1862	South America (Peru, Chile, and Argentina)
Ensis californicus Dall, 1899		Tropical west America
Ensis nitidus (Clessin, 1888)		Tropical west America
Ensis tropicalis Hertlein and Strong, 1955		Tropical west America
Ensis myrae SS Berry, 1953		Tropical west America

Ensis razor shells are distributed along European, African, and American coasts. The fossil record in Europe goes back to the Early Miocene (23.03 - 15.97 mya). Similarly, the north western Atlantic species, *E. directus*, has been described from material which most probably is Miocene as well

(Cosel 2009).

Despite their commercial importance, *Ensis* systematics is still unclear. Some species lack clear morphological autapomorphies (species-specific characters), and therefore they were usually defined based on a combination of characters. The Dutch malacologist Roelof Menno van Urk studied European *Ensis* species from 1964 to 1987 (reviewed in Cosel 2009) and introduced new species and variety names in Van Urk (1964): *E. arcuatus* var. *ensoides*, *E. arcuatus* var. *norvegica*, *E. phaxoides*, and *E. sicula*. In a thorough survey, but using the same approach as Van Urk (shell morphology), those names were synonymised to previous ones by Rudo von Cosel (2009), who recommended the use of molecular tools to clarify systematics. Regrettably, American *Ensis* species (with the exception of *E. directus*) were only briefly reviewed in malacology guides (Huber 2010; Coan and Valentich-Scott 2012). These studies were based on shell morphology only, and lack statistical support.

As a result of the existing difficulties of identifying *Ensis* specimens correctly, several names can be found in the literature to refer to the same species. Even though this has been treated in detail by Cosel (2009), it is worth mentioning that the frequently used names *E. directus* / *E. americanus* (Gould, 1870), and *E. magnus* / *E. arcuatus* (Jeffreys, 1865) are indeed synonyms. The names *E. directus* and *E. magnus* are the senior synonyms, and therefore preferred.

More complicated is the homonymy case of the two *E. minor* species. Since the American species *E. minor* Dall, 1899 is a junior homonym of the European *E. minor* (Chenu, 1843), there is a taxonomic requirement to replace the pre-occupied name *E. minor*, as already recommended by Cosel (1990). In CHAPTER 3.4, we introduced the names *E. coseli coseli* Vierna, 2013 and *E. coseli megistus* Pilsbry and McGinty, 1943. However, these names resulted to be non-code-compliant according to articles 23.3.1, 46.1, and 47.2 of the International Code of Zoological Nomenclature (available at http://www.nhm.ac.uk/hosted-sites/iczn/code/). The correct names (and the ones we will use, except in CHAPTER 3.4) are *E. megistus coseli* Vierna, 2013 and *E. megistus megistus* Pilsbry and McGinty, 1943.

Taking into account the above mentioned taxonomic confusion, it is not unexpected to find mistakes on *Ensis* taxonomy in the scientific literature. For instance, in a comprehensive study on Bristish *Ensis* species, Holme (1954) stated 'they may well be the same species' referring to the American *E. megistus coseli* and the European *E. ensis* (Linné, 1758), despite some morphological differences (e.g. in the shape of the pallial sinus) that are now evident.

In the same way, in the study by Espiñeira et al. (2009), wherein several bivalve species (including

six *Ensis*) were studied, the authors treated *E. directus* and *E. americanus* as different species, even though it is broadly accepted these names are synonyms (Van Urk 1987; Essink 1985; Luczak et al. 1993; Armonies and Reise 1999; Krakau et al. 2006; Cosel 2009).

Currently, there are 13-14 accepted extant *Ensis* species, according to the latest reports (Cosel 2009; Huber 2010; Coan and Valentich-Scott 2012; CHAPTERS 3.3 and 3.4). The species *E. goreensis* (Clessin, 1888) occurs in tropical west Africa and the Cape Verde Islands (Cosel 2009 and references therein). In Europe, there are four extant native *Ensis*, namely *E. magnus*, *E. ensis*, *E. minor*, and *E. siliqua*, in addition to the introduced American species *E. directus*. In Atlantic north America, there are three - four extant taxa: *E. directus*, *E. terranovensis*, *E. megistus coseli*, and *E. megistus megistus.* The south American species *E. macha* occurs in Argentina, Peru, and Chile (Espinoza et al. 2010), and the remaining species, *E. californicus* Dall, 1899, *E. myrae* SS Berry, 1953, *E. nitidus* (Clessin, 1888), and *E. tropicalis* Hertlein and Strong, 1955 are native to tropical west America (Coan and Valentich-Scott 2012) (see Table 1 for a list of extant *Ensis* species, including synonyms).

2.2.3 Biology of *Ensis* species

Ensis are infaunal bivalves that usually inhabit sandy and fine gravel bottoms with limited exposure to wave action. Their abundances can be extremely variable, from just a few individuals in a beach, up to thousands or even millions of them. For example, up to 200 individuals of the species *E. siliqua* and up to 6575 ± 10980 juveniles of *E. directus* were reported to occur in a square meter (Fahy and Gaffney 2001; Dannheim and Rumohr 2012; Jennifer Dannheim personal communication). The maximum size of current adult *Ensis* specimens can exceed 200 mm (see Henderson and Richardson 1994; Cosel 2009). In Galicia, it is uncommon to find *E. magnus* and *E. siliqua* individuals in the same bed (Alicia Pallas personal communication); however, *E. magnus*, *E. ensis*, and *E. siliqua* individuals where sampled at the very same site in Borkum reef (Wadden Sea) (CHAPTER 3.4). In a more recent sampling campaign, *E. directus* has also been found in Borkum reef (Joël Cuperus, personal communication), meaning that several species can co-occur in the same area.

Ensis are gonochoric species, with rare cases of hermaphroditism (see Darriba et al. 2005 and references therein). After fertilisation, the free-swimming planktonic marine larvae trochophores develop into a second larval stage (the veligers). These larvae settle on the seabed and undergo metamorphosis into juveniles known as spat.

Studies on the reproductive cycles of *Ensis* have been performed in commercial species, and have

yielded quite variable results among species and populations. For instance, *E. siliqua* reaches sexual maturity during the first year of age in southern Portugal (Gaspar and Monteiro 1998), after the third year in Wales (Henderson and Richardson 1994), and during the fourth year in eastern Ireland (Fahy and Gaffney 2001).

The species *E. directus* reaches sexual maturity after 1 year (Dannheim and Rumohr 2012 and references therein). Since the calculation of specimen age is not straightforward, some variation could be due to the methodology employed, however differences among populations and species seem to be evident.

Regarding the timing of gametogenesis, it varies among *E. siliqua* populations. For example, in southern Portugal it started in December, and spawning took place only once, between May and July (Gaspar and Monteiro 1998). Similarly, in an eastern Irish population, spawning took place from mid-May to the end of July or early August (Fahy and Gaffney 2001). However, in south western Britain, spawning was reported to be in March-April (Lebour 1938). In a Galician (NW Spain) population, gametogenesis started in November-December, with a unique period of spawning occurring in April-May (Darriba et al. 2005).

On the contrary, in a Galician population of *E. magnus*, gametogenesis started in September - October, and spawning took place several times from December - January, until May - June (Darriba et al. 2004). According to Darriba et al. (2005), differences observed between *E. siliqua* and *E. magnus* from Galicia may have either a genetic or an environmental origin. In Galician populations, settlement took place 19 - 20 days after fertilisation in both species (Martínez 2002; Da Costa et al. 2008).

In a European population of the American species *E. directus*, spawning started in April - May and a second, but weaker, spawning event apparently occurred in August - September (Cardoso et al. 2009). Beukema and Dekker (1995) reported the settlement of European larvae of *E. directus* to be in May - June, although Armonies (1996) registered several temporal pulses of spatfall. According to Cosel (2009) and references therein, in *E. directus*, the duration of the free swimming larval phase ranges between 10 and 27 days, the shortest period (10 days) at a water temperature of 24°C.

Finally, Barón et al. (2004) compared their results on *E. macha* populations from Argentina, with the ones by Avellanal et al. (2002) from Chile. In Argentina, two spawning peaks were detected (in September - November, and in May - June) and mature females were found all year long. In Chile, two peaks of maturation were observed in two out of three analysed populations, and females were found in advanced stages of maturity during the whole year.

2.3 State-of-the-art of genetic studies in razor shells

Evolutionary genetic studies on razor shells are really scarce, probably due to the fact that phylogenies, both at the species and family-levels, as well as species delimitation studies, are still missing. This makes the interpretation of evolutionary genetic results somewhat harder, compared to other animal groups for which phylogenies are well defined, and species boundaries, well established. Recently, the histone genes of the razor shell *Solen marginatus* Pulteney, 1799 were studied in the context of the evolution of these genes in the Protostomia (González-Romero et al. 2009), but our report on the evolution of 5S rDNA in six *Ensis* species (Vierna et al. 2009; see CHAPTER 10) is the first one that deals with the evolution of a gene family in razor shells. Some recent studies have presented new data on the gene order within the mitochondrial genome of *Solen* species and *S. constricta* (Yuan et al. 2012a; Yuan et al. 2012b; Yuan et al. 2012c).

The commercial importance of some razor shell species has probably fuelled population genetic studies. These studies aim to characterise the genetic structure of populations, and to monitor their levels of genetic variation. For example, Arias et al. (2011) assessed the genetic variation of populations of the European species *E. siliqua* (using PCR-RFLPs and RAPDs), in order to 'preserve the natural populations, to promote the rational exploitation of the fishing resources, and to evaluate the possibility of stock enhancement by transplanting individuals from other region'. Varela et al. (2007) developed five primer pairs that amplify polymorphic microsatellite loci in the same razor shell species (*E. siliqua*), and, using these microsatellites, Varela et al. (2012) analysed population variation along the Atlantic coast taking into account the potential effects of the 'Prestige' oil spill. Arias-Pérez et al. (2012) developed five more primer pairs targeting polymorphic microsatellite loci, and analysed population differentiation in the Atlantic. Temporal genetic variation within a population of another European species (*E. magnus*) before and after the 'Prestige' oil spill was also assessed by Varela et al. (2009), using the microsatellite markers reported in the same study.

The microsatellite markers developed by Francisco-Candeira et al. (2007) for the species *S. marginatus*, have not been used so far; however, Hmida et al. (2012) studied three populations of this species from three sampling areas of the Tunisian coasts using isozyme markers. This type of markers were also used to study genetic variability within a population of the Asian species, *Cultellus attenuatus* Dunker, 1862 by Zeng et al. (2010).

Several recent studies have dealt with the Asian species *S. constricta*, which has received much

attention due to its commercial importance. For example, Niu et al. (2007) studied genetic diversity in six populations of this razor shell by sequencing a fragment of the 16S rRNA gene, whereas Niu et al. (2008a) performed a similar study using the COI gene. Niu et al. (2008b) developed eight primer pairs that amplify polymorphic microsatellite loci in this species, and Niu et al. (2009) studied the genetic structure of six geographic populations along the coast of China by means of ISSRs. Wang et al. (2010) studied population variation and differentiation by means of AFLPs markers in China, and 14 new polymorphic microsatellite loci were developed by Jiang et al. (2010). Niu et al. (2012) found significant genetic differentiation among ten Chinese populations using microsatellites and, in their latest report, Niu et al. (2013) sequenced the transcriptome of *S. constricta* and obtained a high number of SNPs and microsatellite-containing sequences.

As for the cytogenetic characterisation of razor shells, prior to this thesis, only four species were characterised: the Chinese *S. constricta* (Wang et al. 1998), and the Atlantic / Mediterranean species *S. marginatus* (Fernández-Tajes et al. 2003), *E. siliqua,* and *E. magnus* (Fernández-Tajes et al. 2008).

Finally, some studies have been published in the last six years on the genetic identification of some razor shell species, mainly *Ensis*. For example, Fernandez-Tajes et al. (2007) established a method to differentiate among four *Ensis* and *S. marginatus* by means of PCR-RFLP of the 5S rDNA region. One year later, Freire et al. (2008) published a method to identify *E. magnus* and *E. siliqua* using the same approach (PCR-RFLPs), but in this case they targeted the ITS1 region. Fernández-Tajes et al. (2010) published a new method, based on the variable size of ITS1 amplicons, intended to identify *E. magnus, E. siliqua, E. directus*, and *E. macha*. Two years later, Fernández-Tajes et al. (2012) established another method, in this case, based on a single microsatellite marker, that attempts to differentiate *E. siliqua* samples from Ireland and the Iberian Peninsula.

In a recent report, Hassan and Kanakaraju (2013) studied the COI sequences of *Solen sarawakensis* Cosel, 2002 individuals along with morphology, from different sampling sites in Malaysia, and their results could support cryptic speciation.

3 GOALS

3. GOALS

The goals of this thesis are:

- To study the diversity and evolution of the 5S ribosomal DNA multigene family in razor shells of the family Pharidae, with a focus on Atlantic *Ensis*, and in other metazoan species whose genomes are sequenced.
- To obtain the karyotypes of the razor shell species *E. directus* (Conrad, 1843) and *E. minor* (Chenu, 1843) and to compare them with those of *E. siliqua* (Linné, 1758), and *E. magnus* Schumacher, 1817.
- To compare current genetic variation in native and introduced populations of the razor shell *E. directus* in order to assess to what extent potential bottlenecks and mass mortality events in the introduced range have impacted diversity; to obtain information about the possible origin of introduced individuals; and to study population structure in the native range of the species.
- To clarify the species status of extant Atlantic *Ensis* morphospecies by means of the analysis of nuclear and mitochondrial genomic regions, and to provide a reliable genetic identification system for these species.

4 RESEARCH ARTICLES

4. RESEARCH ARTICLES

4.1 Evolutionary Studies of 5S Ribosomal DNA

4.1.1 The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae)

Joaquín Vierna, K Thomas Jensen, Andrés Martínez-Lage, Ana M González-Tizón (2011) The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae). Heredity 107:127-142.

Bibliometrics 2012 JCR Science Edition

Impact factor: 4.110

Ecology: Q1

Evolutionary Biology: Q2

Genetics & Heredity: Q1

Heredity (2011) 107, 127–142

www.nature.com/hdy

ORIGINAL ARTICLE

The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae)

J Vierna[1], KT Jensen[2], A Martínez-Lage[1] and AM González-Tizón[1]
[1]*Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, La Coruña, Spain and* [2]*Marine Ecology, Department of Biological Sciences, Aarhus University, Ole Worms Allé 1, Aarhus C, Denmark*

The linkage between 5S ribosomal DNA and other multigene families has been detected in many eukaryote lineages, but whether it provides any selective advantage remains unclear. In this work, we report the occurrence of linked units of 5S ribosomal DNA (5S rDNA) and U1 small nuclear DNA (U1 snDNA) in 10 razor shell species (Mollusca: Bivalvia: Pharidae) from four different genera. We obtained several clones containing partial or complete repeats of both multigene families in which both types of genes displayed the same orientation. We provide a comprehensive collection of razor shell 5S rDNA clones, both with linked and nonlinked organisation, and the first bivalve U1 snDNA sequences. We predicted the secondary structures and characterised the upstream and downstream conserved elements, including a region at −25 nucleotides from both 5S rDNA and U1 snDNA transcription start sites. The analysis of 5S rDNA showed that some nontranscribed spacers (NTSs) are more closely related to NTSs from other species (and genera) than to NTSs from the species they were retrieved from, suggesting birth-and-death evolution and ancestral polymorphism. Nucleotide conservation within the functional regions suggests the involvement of purifying selection, unequal crossing-overs and gene conversions. Taking into account this and other studies, we discuss the possible mechanisms by which both multigene families could have become linked in the Pharidae lineage. The reason why 5S rDNA is often found linked to other multigene families seems to be the result of stochastic processes within genomes in which its high copy number is determinant.
Heredity (2011) **107,** 127–142; doi:10.1038/hdy.2010.174; published online 2 March 2011

Keywords: birth-and-death evolution; regulatory regions; 5S ribosomal RNA; U1 small nuclear RNA; linkage; *Ensis*

Introduction

The 5S ribosomal RNA molecule (5S rRNA) is a component of the large subunit of ribosomes, encoded by the 5S ribosomal DNA (5S rDNA) and transcribed by RNA polymerase III. The eukaryote 5S rDNA is a multigene family, typically composed of hundreds of repeats of an approximately 120 nucleotides (nts) RNA coding region (hereafter, 5S) and an intergenic spacer (IGS) usually referred to as nontranscribed spacer (NTS). The first nts downstream the 5S are transcribed as part of the primary RNA and deleted during RNA maturation (Sharp *et al*., 1984; Sharp and Garcia, 1988), but they are considered as part of the NTS.

The 5S rDNA is characterised by a flexible organisation, as it has been found in clusters composed of similar or divergent tandemly arranged repeats (differences mainly occur within the NTS; for example, Shippen-Lentz and Vezza, 1988), and in clusters of 5S rDNA repeats tandemly linked to other multigene families (for example, Cross and Rebordinos, 2005; Freire *et al*., 2010; Cabral-de-Mello *et al*., 2010). A dispersed organisation of 5S rDNA has also been reported (Morzycka-Wroblewska *et al*., 1985 and references therein), and some species were found to have more than one type of organisation within the genome (Little and Braaten, 1989).

The 5S rDNA multigene family was thought to be characterised by low levels of intragenomic divergence in virtually all species because of the concerted evolution of ribosomal multigene families (see Eickbush and Eickbush, 2007 for a review). Nevertheless, the ocurrence of divergent variants of 5S rDNA within a genome has been described in animals, plants and fungi (for example, Fernandez *et al*., 2005; Rooney and Ward, 2005; Caradonna *et al*., 2007), and in some cases, differences in the RNA coding regions were found to correspond to tissue-specific variants (Peterson *et al*., 1980). Therefore, recent studies have pointed out to a more complex evolutionary scenario in which birth-and-death processes generate new 5S rDNA variants that may be homogenised by unequal crossing-overs and gene conversions. For instance, in *Ensis* razor shells (Schumacher, 1817), the long-term evolution of 5S rDNA was found to be driven by birth-and-death processes and selection, and it was suggested that homogenising mechanisms were also taking part within each variant in each species (Vierna *et al*., 2009). Later on, it was proposed that the levels of intragenomic divergence—much higher within the 5S rDNA than within the major ribosomal genes—were due to the more flexible organisation of 5S rDNA, meaning that homogenisation processes were more efficient within the array(s) of major ribosomal genes, as they may occur in a smaller number. The long-term

Correspondence: J Vierna or AM González-Tizón, Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, A Fraga, 10, La Coruña E-15008, Spain.
E-mails: jvierna@udc.es; joaquinvierna@gmail.com or hakuna@udc.es
Received 12 July 2010; revised 18 October 2010; accepted 8 November 2010; published online 2 March 2011

evolution of both rDNA regions was then proposed to be driven by a mixed process of concerted evolution, birth-and-death evolution and purifying selection, as described by Nei and Rooney (2005) (Vierna *et al*., 2010).

Most eukaryotic genes are transcribed into precursor messenger RNAs that must undergo splicing, an essential step of gene expression. During precursor messenger RNA splicing, introns are removed from the precursor messenger RNA and exons are ligated together to form mRNA (Will and Lührmann, 2005). Splicing is performed by the spliceosomes, ribonucleoprotein complexes consisting of small nuclear RNAs and several proteins. The U1 small nuclear RNA molecule is a component of the major spliceosome, essential for the interaction with the 5′ splice site of introns (Zhuang and Weiner, 1986). This molecule is encoded by the U1 small nuclear DNA (U1 snDNA), which consists of an RNA coding region (hereafter, U1) and an IGS (when it is organised in tandem repeats). U1 snDNA, transcribed by RNA polymerase II, is a multigene family with a variable number of repeats in each genome (around tens of repeats in the metazoan species studied by Marz *et al*., 2008). Although not much information is available about the organisation of U1 snDNA, it was found to be linked to other multigene families, such as 5S rDNA (Pelliccia *et al*., 2001), other spliceosomal snDNA families (Marz *et al*., 2008) and organised in the same array together with 5S rDNA repeats and other spliceosomal snDNA (Manchado *et al*., 2006). In general, however, clustered copies of distinct or the same small nuclear RNA coding genes are not common in metazoan genomes (Marz *et al*., 2008).

The evolution of spliceosomal snDNA has been recently studied in two different surveys, covering insect species (Mount *et al*., 2007) and several other metazoan groups (Marz *et al*., 2008), and appears not to be a simple issue. In insects, it is governed by several concurrent forces, namely purifying selection, unequal crossing-overs, gene conversions and birth-and-death processes (Mount *et al*., 2007). Distinguishable U1 snDNA paralogs differentially expressed throughout development have been described in some species (for example, Lo and Mount, 1990), but the snDNA paralog groups seem not to be stable over a long evolutionary time, although they appear independently in several clades (Marz *et al*., 2008).

The linkage between 5S rDNA and U1 snDNA has only been reported in one crustacean (Pelliccia *et al*., 2001) and in one fish (Manchado *et al*., 2006). In this survey, we report linked units of 5S rDNA and U1 snDNA in 10 razor shell species (Mollusca: Bivalvia: Pharidae) from four different genera. We obtained new data about the genomic organisation of both multigene families in these animals, and studied the genesis and evolution of the 5S rDNA–U1 snDNA linked units. Using the *Ensis* sequences available from DDBJ/EMBL/GenBank and the new sequences obtained, we provide a comprehensive collection of razor shell 5S rDNA variants, including their secondary structures and the characterisation of putative pseudogenes. We also report the first Bivalvia U1 snDNA sequences, including their predicted secondary structures. Finally, several putative regulatory regions of both multigene families were studied in detail.

Materials and methods

Animals

We selected 11 species belonging to family Pharidae (Adams and Adams, 1858; Mollusca: Bivalvia, Table 1). Though a greater sampling effort was made on genus *Ensis*, we tried to represent the whole family by selecting species from its different subtaxons. Thus, from subfamily Cultellinae (Davies, 1935), we studied eight *Ensis* species and one *Ensiculus* (Adams, 1860). The species *Siliqua patula* (Dixon, 1789) was also included in the analysis, as genus *Siliqua* (Mühlfeld, 1811) may represent a separate subfamily from the Cultellinae (see Cosel, 1993). From the other subfamily, Pharinae (Adams and Adams, 1858), we took into consideration the species *Pharus legumen* (Linné, 1758). Two homonymous species, *Ensis minor* (Chenu, 1843) and *E. minor* (Dall, 1899) were studied in this survey, and hereafter they will be referred to as *E. minor* (Chenu) and *E. minor* (Dall). All taxon names follow Cosel (1993) and Cosel (2009), when applicable. Razor shells were provided by several colleagues and preserved in 100% ethanol until species identification, except the *Ensiculus cultellus* (Linné, 1758) sample that consisted of an ethanol-preserved piece of muscle tissue, and *Ensis goreensis* (Clessin, 1888) from which only dried tissue was available.

DNA extraction, PCR, cloning and sequencing

DNA extractions were done from muscle tissue using the NucleoSpin Tissue kit (Macherey-Nagel, North Rhine-Westphalia, Germany). Using the primers 5S-Univ-F and 5S-Univ-R (Vierna *et al*., 2009), we serendipitously amplified complete U1 snDNA sequences flanked by two partial 5S rDNA repeats in the species *Ensis magnus* (Schumacher, 1817) and *P. legumen*. From these sequences, different primer pairs annealing at the 5S and U1 regions of razor shells were designed using GeneFisher (Giegerich *et al*., 1996) (Table 2). PCR reactions were conducted in a final volume of 20 μl using the 2 × Taq Master Mix RED (VWR/Ampliqon, Skovlunde, Denmark), applying the following conditions: an initial denaturation step at 94 °C for 3 min followed by 40 cycles of denaturation at 94 °C for 20 s, annealing at the temperatures indicated in Table 2 for 20 s, extension at 72 °C for 1 min, and a final extension at 72 °C for 5 min. Amplification products were run on 1% agarose gels, stained with a 0.5 μg/ml solution of ethidium bromide, and imaged under UV light. They were cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA). A subset of transformant colonies from each cloning reaction was analysed by PCR in order to check the insert size. From each PCR, we selected one clone per species when only one band was retrieved (that is, in all cases except in one of the PCRs of *Ensis macha* (Molina, 1782) and *E. cultellus* individuals, in which case we obtained two slightly different bands, so two clones were sequenced). Sequencing was performed at Macrogen (Seoul, South Korea) using both T3 and T7 primers (forward and reverse) included in the cloning kit.

Bioinformatic analyses

Electropherograms were inspected in BioEdit 7.0.9.0 (Hall, 1999). The Blast 2 sequences tool (available at www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi) was

Table 1 DNA sequences studied and specimen details

Species		*sbf.*	*Identification*	*Museum code*	*Sampling site*	*Primer pair names and accession numbers*			
						5S-Univ	*5S-U1*	*U1-5S*	*U1-U1*
Ensis magnus	Schumacher, 1817	Cul.	Cosel, 2009	MNHN 40042	Bonden, Sweden	FN908876*a*	FN908883*a*	FN908894*a*	FN908904*a*
E. magnus	Schumacher, 1817	Cul.	Cosel, 2009		Ortigueira, Spain	FM201454-56*b*			
E. siliqua	(Linné, 1758)	Cul.	Cosel, 2009	MNHN 40047	Vigo, Spain	FM201457-62*b*, FM211689*b*	FN908890*a*	FN908900*a*	FN908908*a*
E. ensis	(Linné, 1758)	Cul.	Cosel, 2009	MNHN 40044	La Capte, France	FM211690-91*b*	FN908885*a*	FN908896*a*	FN908905*a*
E. gorezensis	(Clessin, 1888)	Cul.	Cosel, 2009	MNHN 17948	Gorée, Senegal	FM211692*b*			
E. minor	(Chenu, 1843)	Cul.	Cosel, 2009	MNHN 40045	La Capte, France		FN908886*a*	FN908897*a*	FN908906*a*
E. directus	(Conrad, 1843)	Cul.	Cosel, 2009	MNHN 40049	Long Pond, Canada		FN908884*a*	FN908895*a*	
E. directus	(Conrad, 1843)	Cul.	Cosel, 2009		Various localities, Denmark	AM904878-933*b*			
E. macha	(Molina, 1782)	Cul.	Cosel, 2009	MNHN IM-2009-8446	Puerto Lobos, Argentina		FN908887*a*, FN908888*a*	FN908898*a*	FN908907*a*
E. macha	(Molina, 1782)	Cul.	Cosel, 2009	MNHN 40048	Playa Dichato, Chile	FM201452*b*			
E. macha	(Molina, 1782)	Cul.			Concepción, Chile	AM940998-1009*b*/*c*, AM906171-80*b*/*c*, AM906203-8*b*/*c*			
E minor	Dall, 1899	Cul.	Cosel, 2009	MNHN IM–2009–8447	Christmas Bay, USA		FN908889*a*	FN908899*a*	
Ensiculus cultellus	(Linné, 1758)	Cul.	John Taylor	BMNH 20070223	Moreton Bay, Australia		FN908881*a*, FN908882*a*	FN908893*a*	FN908903*a*
Siliqua patula	(Dixon, 1789)	Cul.	Dan Ayres	MNHN IM–2009–8448	Ocean City, USA		FN908892*a*	FN908902*a*	FN908910*a*
Pharus legumen	Linné, 1758	Phar.	Cosel, 2009	MNHN 40051	Bandol, France	FN908877-80*a*	FN908891*a*	FN908901*a*	FN908909*a*

Abbreviations: *a*, new sequences; *b*, sequences previously studied by Vierna *et al*. (2009); *c*, sequences previously studied by Fernández-Tajes and Méndez, (2009); Cul., Cultellinae; Phar., Pharinae; sbf., subfamily.

Table 2 Primer pairs used in this survey

Sequence/reference		*T*	*a.r.*	*a.p.*
5S-Univ-F	Vierna *et al*. (2009)	50 °C	5S	13–32
5S-Univ-R	Vierna *et al*. (2009)	50 °C	5S	36–55
5S-U1-F	5′ GTCTACGGCCATATCACGTT	61 °C	5S	1–20
5S-U1-R	5′ GTTAGCGCGAACGCAGVC	61 °C	U1	142–159
U1-5S-F	5′ VCTGCGTTCGCGCTAVCC	65 °C	U1	143–160
U1-5S-R	5′ GGTATTCCCAGGCGGTCAC	65 °C	5S	87–105
U1-U1-F	5′ GCAATGGAAGGGCCTCCTCCT	61 °C	U1	49–69
U1-U1-R	5′ TTCGGTTGGGCTGATGCCTG	61 °C	U1	72–91

Abbreviations: a.p., annealing position within each RNA coding region; a.r., annealing region; T, annealing temperature; U1, U1 small nuclear RNA coding region; 5S, 5S ribosomal RNA coding region.

used to compare the ends of both the forward and reverse sequences obtained from each clone, which were subsequently overlapped by hand. Sequences obtained were subjected to a sequence-similarity search against the DDBJ/EMBL/GenBank nucleotide collection databases using the blastn algorithm. Sequences similar to other 5S, U1 and their intergenic spacers were deposited in the DDBJ/EMBL/GenBank databases under the accession numbers specified in Table 1. The pair-wise comparisons were also performed in the Blast 2 sequences tool and multiple sequence alignments were carried out in ClustalW 2.0 (Larkin *et al*., 2007), and manually adjusted for local optimisation in MEGA 4.0.2 (Tamura *et al*., 2007). The number of polymorphic sites was retrieved from DnaSP 5.10.0 (Librado and Rozas, 2009). Lengths and p-distances were obtained from MEGA 4.0.2 (Tamura *et al*., 2007). In p-distance calculation, gaps were not considered, and 1000 bootstrap replicates were performed for the estimation of standard errors.

In order to search for putative regulatory conserved elements, sequences upstream and downstream the 5S and U1 regions were analysed. Searches were performed considering the first 100 nt upstream and downstream the RNA coding regions. In the case of U1 upstream analyses, two sequences from the gastropod molluscs *Aplysia californica* and *Lottia gigantea* (provided by Manja Marz, Philipps-Universität, Marburg, Germany) were selected and included in the analyses. Conserved motifs were identified by MEME (Bailey and Elkan, 1994) and they were manually compared with published regulatory elements.

5S and U1 sequences were folded in RNAstructure 5.02 (Reuter and Mathews, 2010) at 15 °C, and we used the efn2 function (Mathews *et al*., 1999) to recalculate the ΔG values. The consensus secondary structures were obtained from the RNAalifold webserver (Hofacker, 2003).

We used PALM (Chen *et al*., 2009) to select nucleotide substitution models and to infer maximum likelihood phylogenies. The best-fit model of nucleotide substitution was directly selected using Modeltest 3.7 (Posada and Crandall, 1998), applying the Akaike information criterion. Phylogenies were constructed by PALM using PhyML (Guindon and Gascuel, 2003). Starting trees were obtained by the BioNJ algorithm (Gascuel, 1997) and

gaps were treated as unknown characters. The number of substitution rate categories employed was eight, and the bootstrap test (Felsestein, 1985) was used to estimate node support (1000 replicates). Maximum parsimony phylogenies were obtained from PAUP*4.0b10 (Swofford, 2002) as detailed in Vierna *et al*. (2010). Following Marz *et al*. (2008), we calculated phylogenetic networks in addition to phylogenetic trees, using the neighbour-net algorithm (Bryant and Moulton, 2004), implemented as part of the SplitsTree4 package (Huson and Bryant, 2006).

Different gene tandem arrangements were drawn using pDRAW32 (AcaClone sofware, http://www.acaclone.com/) and we edited all phylogenetic trees in FigTree 1.2.2 (Andrew Rambaut, http://tree.bio.ed.ac.uk/software/figtree/).

Results

Sequence characterisation

The identification of 5S, U1 and spacer sequences was performed by comparing them against the DDBJ/EMBL/GenBank nucleotide collection databases, as explained above. For the sake of clearness, all spacer sequences downstream a 5S will be referred to as NTS, and all spacers downstream a U1, as IGS. All complete 5S sequences were 120 nts and NTS ranged between 283 and 986 nts. All complete U1 sequences were 164 nts except the ones obtained from *S. patula*, that had a nucleotide insertion at position 37. IGS ranged between 222 and 422 nts. The DDBJ/EMBL/GenBank accession numbers of the sequences studied are listed in Table 1.

Average GC contents were 55.1% for the 5S region, 54.8% for the U1 region, 38.8% for the NTS and 41.9% for the IGS. The number of polymorphic sites in the RNA coding regions was $S = 32$ for the 5S region and $S = 20$ for the U1 region.

Hereafter, clones containing partial or complete repeats of both multigene families will be referred to as mixed clones.

Alignments

An initial alignment of the NTS region showed that the NTSs of razor shells were highly divergent, so sequences had to be grouped separately, according to their similarity. After performing several combinations, we divided the NTSs into seven supergroups and 17 groups. Each supergroup was named using a Roman numeral and each group was denoted by a Greek letter following Vierna *et al*. (2009). Supergroups and groups contained sequences belonging to one or more species. Similarily, IGS sequences were divided into two groups, one containing all *Ensis* and *Ensiculus* and the other one containing *Pharus* and *Siliqua* IGSs. The species composition, lengths and mean *P*-distances for each spacer group and supergroup were recorded in Table 3.

Let's now consider only the spacer sequences from mixed clones. We were able to align all IGSs from *Ensis*, *Ensiculus*, *Pharus* and *Siliqua* individuals, but the divergence among them was evident; however, the last part of the alignment (containing the upstream region of the next 5S repeat) revealed a more conserved region. Quite the opposite, the analysis of the NTSs from mixed clones (upstream U1 sequences) revealed that these spacers were less conserved than the IGSs and could not be aligned at once. In this case, we were able to align all *Ensis* sequences (from supergroup II), except an NTS from the species *E. macha* (from supergroup V). The NTSs from the species *P. legumen* and *S. patula*, belonging to supergroup IV, could also be aligned together. However, *E. cultellus* NTSs could not be aligned to *Ensis*, *P. legumen* or *S. patula* sequences.

In the alignment of *Ensis* U1–U1 clones (Supplementary File S1), all *Ensis* IGSs displayed a region of similarity with δ- and γ-NTSs, from the species *E. directus* (Conrad, 1843). This region was located at the end of the IGS (just upstream the 5S region) and resembled the last portion of δ- and γ-NTSs. Downstream this 5S region, in the NTS, we found another region of similarity with δ- and γ-NTSs, and downstream of it there was a fragment resembling a 5S (probably an old pseudogenised copy). Even though this pattern was only found in *Ensis* species, the first portion of the alignment that corresponded to the U1–IGS–5S sequence (positions 1 to 427, Supplementary file S1), could be aligned to *E. cultellus* clones, and with more difficulties, to *P. legumen* and *S. patula* ones (as explained above).

Upstream elements

A conserved region was identified at −25 nts from both the 5S rDNA and U1 snDNA transcription start sites (Supplementary file S2) and named −25 region. It was a TATA-like motif in the 5S upstream sequences (Supplementary file S3a), and upstream the U1 region (Supplementary file S3b), it was an A/G-rich motif: AAAAG in *Ensis* and *E. cultellus*, GGGGA in gastropods, AAATG in *P. legumen* and GTAAG upstream *S. patula* putative-pseudogenised U1 sequences (see U1 predicted secondary structures). Another motif (AAAGC, Supplementary file S2) was identified just upstream the U1 snDNA transcription start site, identical to the one found in *Drosophila melanogaster* (Lo and Mount, 1990) and in other organisms (see Discussion), but it only occurred in some of the razor shell sequences. Finally, a less conserved region was found upstream the −25 region in U1 snDNA upstream sequences (Supplementary file S2), centred at −44 nts.

Although it was not possible to align all *Ensis* NTSs at once, we were able to align the 100 nt upstream the transcription start site of 5S rDNA of *Ensis* species. These stretches were the last part of either NTS or IGS sequences. We failed to include the other Pharidae species in this alignment, as sequences were not conserved among genera.

Internal regulatory regions

5S internal control regions (ICR I to IV) were compared with those described in *D. melanogaster* (Sharp and Garcia, 1988). As some ICRs coincided with the primer-annealing regions, some sequences were excluded from the comparisons, and sequences amplified with the 5S-Univ primers (Table 2) were only included in the ICR IV analysis. Results were as follows: 12/16 matches within ICR I (positions 3–18); 7/8 matches within ICR II (positions 37–44); 11/14 matches within ICR III (positions 48–61); and 14/21 matches within the ICR IV region (positions 78–98). The degree of conservation of these elements within razor shells was of 14/16, 8/8, 13/14

Table 3 Intergenic spacer groups and supergroups

NTS group	*Species*	*Clade*	N	*Length*	*Mean P-distance*
Supergroup I			72	286–329	0.135 ± 0.010
α	*Ensis directus*	A	41	321–329	0.011 ± 0.003
β	*Ensis macha*	A	18	314–318	0.019 ± 0.004
ζ	*Ensis magnus, E. siliqua, E. ensis, E. goreensis*	E	13	286–315	0.042 ± 0.006
Supergroup II			28	407–965	0.240 ± 0.012
γ	*Ensis directus*	A	11	444–654	0.023 ± 0.004
δ	*Ensis directus*	A	4	407	0.002 ± 0.002
η*	*Ensis magnus, E. siliqua, E. ensis, E. minor* (Chenu)	E	9	893–965	0.046 ± 0.004
θ*	*Ensis directus, E. macha, E. minor* (Dall)	A	4	926–960	0.127 ± 0.008
Supergroup III			14	405–620	0.141 ± 0.009
ε 1	*Ensis macha*	A	6	603	0.011 ± 0.003
ε 2	*Ensis macha*	A	5	618–620	0.004 ± 0.002
ξ*	*Ensiculus cultellus*		3	405	0.010 ± 0.004
Supergroup IV			8	355–550	0.241 ± 0.013
μ	*Pharus legumen*		3	548–550	0.005 ± 0.002
*o**	*Siliqua patula*		2	355	0
λ*	*Pharus legumen*		3	419–420	0
Supergroup V					
ι*	*Ensis macha*	A	1	776	
Supergroup VI			2	209–369	0.077 ± 0.018
π*	*Siliqua patula*		1	369	
ρ*	*Siliqua patula*		1	209	
Supergroup VII			2	283–332	0.366 ± 0.029
κ	*Pharus legumen*		1	332	
ν*	*Ensiculus cultellus*		1	283	
IGS group	*Species*		n	*Length*	*Mean P-distance*
Supergroup Ensis–Ensiculus			15	222–422	0.203 ± 0.015
Ensis spp.	*Ensis directus, E. macha, E. minor* (Dall, 1899) *Ensis magnus, E. siliqua, E. ensis, E. minor* (Chenu, 1843)		13	225–231	0.177 ± 0.014
Ensiculus cultellus	*Ensiculus cultellus*		2	421–422	0.002 ± 0.002
Supergroup Pharus–Siliqua			5	236–342	0.193 ± 0.017
Siliqua patula	*Siliqua patula*		2	236	0.064 ± 0.015
Pharus legumen	*Pharus legumen*		3	342	0.007 ± 0.004

Abbreviations: A, American clade; E, European clade (*Ensis* phylogenetic clades according to Vierna *et al*. (unpublished data)); IGS, intergenic spacer (downstream a U1 small nuclear RNA coding region); n, sample size, NTS, nontranscribed spacer (intergenic spacer downstream a 5S ribosomal RNA coding region).
Asterisks (*) indicate nontranscribed spacers linked to U1 small nuclear DNA;

and 15/21 matches, respectively. Similarly, positions 50–61 (Box A), 80–89 (Box C) and 62–79 (intermediate sequence) were compared with those described by Pieler *et al*. (1987) in *Xenopus laevis*, obtaining 6/12, 6/10 and 12/18 matches. Within razor shells, the matches obtained were 9/12, 7/10 and 14/18.

Six U1 internal regions that appear to be conserved accross metazoa (Zhuang and Weiner, 1986; Marz *et al*., 2008) were analysed in all razor shell sequences. They were compared with the two gastropod sequences (see above), the ones from the insect *D. melanogaster* (Lo and Mount, 1990), and those from crustaceans *Asellus aquaticus* and *Proasellus coxalis* (Barzotti *et al*., 2003). Considering as a reference the *E. magnus* U1 sequence (see U1 predicted secondary structures), they correspond to the following positions: the 5′ end (includes the 5′ splice site, Zhuang and Weiner, 1986 and references therein); 28–33 (within the U1–70 K protein binding site, Query *et al*., 1989); the stem-loop II positions 53–55, 65–72 and 84–86 (U1-A protein binding region, Scherly *et al*., 1989); and positions 124–132 (include the Sm protein binding region, named 'domain A' by Branlant *et al*., 1982). The most conserved region was the 5′ end (11 nt) that was identical in all sequences. Positions 28–33 were identical in all sequences, but *L. gigantea* had an additional G inserted between the first and the second nt. Positions 65–72 were also identical, except in the last nt. Finally, the 124–132 region was also conserved with the exception of the sixth and last nt. The remaining two regions were conserved at positions 54–55 and 84–85 in all molluscs and arthropods.

Termination signals

One or more TTTT stretches (required for 5S rDNA transcription termination, Bogenhagen and Brown, 1981; Huang and Maraia, 2001; Richard and Manley, 2009) occurred within the first 20 nt of all NTSs, except for

those belonging to the δ-group, for which the first perfect TTTT was located at positions 96–99. All NTSs except those from the α-group had a TTT motif within the first six nts, and 124/125 sequences had a T residue in the first position.

The analysis of the first portion of the IGS revealed that a TAAAA motif occurred in all *Ensis* species and *E. cultellus*, contiguous to the 3′ end of the U1. Sequences from *S. patula* had a TCCAT and those from *P. legumen*, an ATATA motif. All sequences displayed between two and four AAT stretches within the first 88 sites downstream the U1 region. However, no other evidence of conserved regions that could be involved in the formation of the 3′ end (as the 3′ box, Hernandez, 1985) were found. The first 50 sites downstream the U1 region were very rich (44.5%) in adenines.

Genomic organisation

Mixed clones of 5S rDNA and U1 snDNA were retrieved from all species analysed, except from *E. goreensis* (the quality of the extracted DNA was very low) and both multigene families displayed the same orientation. Both U1–5S and 5S–U1 primer pairs yielded PCR products, and amplifications using U1–U1 primers were successful in eight of them (see primer details in Table 2). Tandemly-arranged 5S rDNA repeats (a partial 5S, an NTS and a partial 5S) were retrieved from *P. legumen* and six *Ensis* species: *E. goreensis*, *E. magnus*, *E. siliqua* (Linné, 1758), *E. ensis* (Linné, 1758), *E. directus* and *E. macha*; two 5S rDNA repeats flanked by U1 snDNA were sampled from the species *S. patula* and *E. cultellus* (Figures 1a and b) and tandemly-arranged U1 snDNA repeats were not found in any of the species studied, as clones obtained with the U1–U1 primers always had one or two 5S rDNA repeats in between. The sequence analysis of clone ends permitted us to determine which clones could be overlapped, assuming that identical spacer sequences retrieved from different clones from the same individual were, in fact, the same copy. Therefore, a sequence of 2217 nts was obtained from *P. legumen* clones (Figure 1c). In all *Ensis* species, the organisation of mixed clones was very similar and consisted of two partial U1 snDNA repeats flanking one complete 5S rDNA repeat and/or vice-versa (for example, Figure 1d).

One of the two 5S–U1 clones from the species *E. macha* (Supplementary file S4) was different from all other *Ensis* clones. It consisted of a partial 5S followed by a divergent NTS (from group ι, supergroup V). This NTS contained a region of similarity with ε-2 NTSs (from the species *E. macha*), a 50 nts truncated 5S copy, 95 nts very similar to the previous ε-2-similar region and a region that matched to a sequence associated to a *Taenia solium* spliced leader and spliced leader mini-exon (from Brehm *et al.*, 2002) that appeared to be a silent DNA (Klaus Brehm, personal communication). At the end of the clone, we found the type A U1 sequence (see U1 predicted secondary structures) that was somewhat divergent, with respect to the other *Ensis* U1s (see Phylogenetic trees and networks).

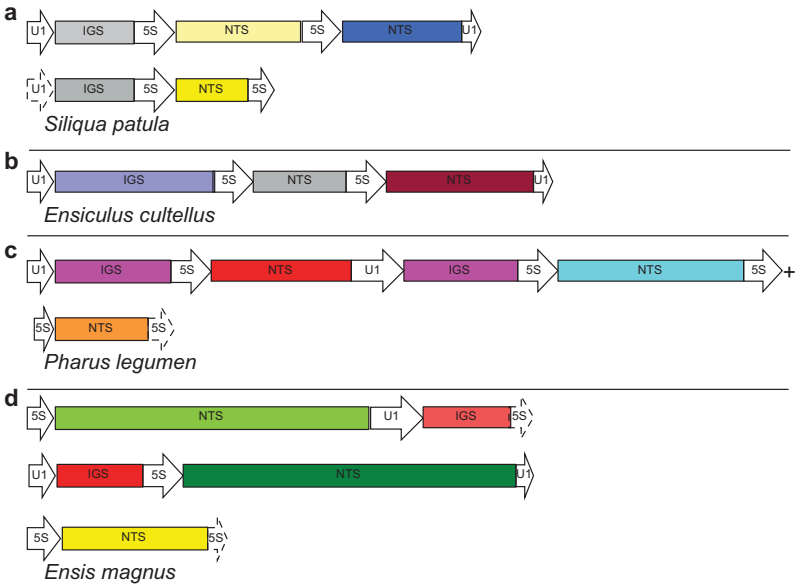


Figure 1 Different 5S ribosomal DNA and U1 small nuclear DNA tandem arrangements sampled from razor shell species. For each species, drawings were constructed using sequences retrieved from the same individual. Drawings are done to scale (except dash lined boxes). (**a**) Grey IGSs are very similar (identities = 90%, gaps = 2%, E value = 1×10^{-93}); yellow NTSs are similar but the darker one has a deletion of 160 nts (identities = 90% in both aligned regions, gaps = 4% in the first region and gaps = 1% in the second region, E value = 4×10^{-58}) . Blue and yellow NTSs are very divergent and could not be aligned. Blue, *o*-NTS; light yellow, π-NTS; dark yellow, ρ-NTS. (**b**) NTSs are very divergent and could not be aligned. Grey, ν-NTS; brown, ξ-NTS. (**c**) Both IGS (same colour) are identical. The three NTS are very divergent and could not be aligned. Red, λ-NTS; light blue, μ-NTS; orange, κ-NTS. (**d**) Green NTSs (both η) are very similar (identities = 82%, gaps = 9%, E value = 0), red IGSs are also very similar (identities = 92%, no gaps, E value = 1×10^{-93}). Yellow NTS corresponds to ζ-group. 5S, 5S ribosomal RNA coding region; U1, U1 small nuclear RNA coding region; NTS, nontranscribed spacer; IGS, intergenic spacer. (+) Reconstructed by overlapping clones from the same individual (see main text).

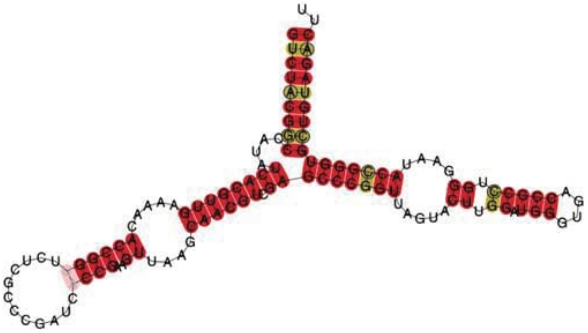


Figure 2 Predicted consensus secondary structure of razor shell 5S ribosomal RNA. Helices are named with Roman numerals and letters correspond to loops, following Barciszewska *et al*. (2000). Red indicates one type of base pair and ochre indicates two types of base pairs. Pale colours indicate pairs that cannot be formed by all sequences.

5S predicted secondary structures

Fourteen different sequences that contained a complete 5S after excluding the primer-annealing regions were considered for the secondary structure prediction. The predicted structures of 11 sequences (Supplementary file S5 a-k) were consistent with the general secondary structure of 5S rRNA (Barciszewska *et al*., 2000). The presence of two fixed thymines (uracils in the RNA molecule) in positions 80 and 96 caused the formation of an additional mini-loop in helix IV (Figure 2, Supplementary file S5).

The other three sequences (from the species *E. ensis*, *E. siliqua* and *E. directus*) may be pseudogenised sequences, because they displayed abnormal secondary structures (Supplementary file S5 l-n). The ΔG values (Table 4) of these putative pseudogenes were the most positive ones, indicating that the structures are less stable.

Compensatory changes C–G→A–T occurred at paired positions 8–111 in the *E. cultellus* sequence (Supplementary file S5 i). A C→T change in position 45 (Supplementary file S5 j and k) made helix III to be one nt smaller in both *P. legumen* predicted structures.

U1 predicted secondary structures

After excluding the primer-annealing regions, 30 razor shell U1 sequences ranging between 48 and 164 nts were obtained. Two of them, from the species *E. magnus* and *P. legumen*, were complete sequences. The species *E. ensis* and *E. siliqua* had identical sequences to *E. magnus*. Complete U1s were build up by overlapping sequences obtained with primer pairs U1–U1 and 5S–U1 (Table 2) from *E. minor* (Chenu), *E. cultellus* and *S. patula* individuals. Three different clones containing partial U1s were retrieved from the *E. macha* individual (two 5S–U1 and one U1–U1) and two different U1 types, named A and B, were characterised. *E. macha* type A U1 was retrieved from a 5S-U1 clone and lacked 23 nt from its 3′ end, which were completed for secondary structure prediction with the corresponding 3′ end of *E. magnus* U1. *E. macha* type B U1 was built up by overlapping the sequence obtained with primers U1–U1 with the sequence obtained from the other 5S–U1 clone. The *E. directus* and *E. minor* (Dall) sequences were also completed in order to predict their secondary structures. Complete U1s from the gastropod molluscs *A. californica* and *L. gigantea* were included, as explained above, in the analyses.

U1 predicted secondary structures were in agreement with previously proposed ones (for example, that of *D. melanogaster*, Lo and Mount, 1990; and consensus structures for several metozoan groups, Marz *et al*., 2008) (Figure 3; Supplementary file S6). The secondary structure of stem-loop IV was conserved in razor shells and gastropods and consisted of one hairpin loop, one internal loop, two stems and a central nonpaired region containing the 5′ splice site (Zhuang and Weiner, 1986 and references therein) and the Sm proteins binding region ('domain A', Branlant *et al*., 1982). The predicted secondary structure of stem-loop I was similar in all U1s, except *E. macha* type A U1 and *L. gigantea* sequences (Supplementary file S6 c and i), whose internal loops were 2–3 nts bigger. *E. macha* type B U1, *E. directus*, *A. californica* and *E. minor* (Dall) sequences had two internal loops in stem-loop III (Supplementary file S6 d, e, h, and j), whereas the rest of them had three. Stem-loop II had two internal and one hairpin loop in all structures, except in *L. gigantea* and *E. minor* (Dall) (Supplementary file S6 i and j), which lacked an internal one. However, this sequence and the one from *S. patula* were considered as putative pseudogenised copies because of their abnormal predicted secondary structures and their ΔG values (Supplementary file S6 j and k; Table 4).

Phylogenetic trees and networks

Phylogenetic trees were constructed both under maximum likelihood and maximum parsimony criteria and the resulting tree topologies were consistent in all cases. All trees shown were constructed with no outgroups.

The phylogenetic trees and the network performed with the 5S sequences of razor shells did not show a clear clustering by species. However, they also failed to show a well-supported clustering by 5S variants (Supplementary file S7).

We were able to align the upstream region of *Ensis* 5S rDNA, and the phylogenies performed showed that 5S rDNA can be divided into four different groups according to their upstream sequences (see Figure 4).

Because of the impossibility of aligning all NTSs at once due to their high degree of divergence, NTS phylogenies were performed considering each one of

Table 4 ΔG values calculated at 15 °C using the efn2 function for each predicted secondary structure

ΔG (*kcal mol*$^{-1}$)	*Species*	*NTS group*	*Linked to U1 snDNA?*	*Complete sequence?*	*m.s. (nts)*
5S rRNA					
−46.5	*Ensis ensis*	ζ	No	Yes	120
−47.6	*Ensis siliqua*	η	Yes	Yes	120
−49.6	*Ensis directus*	α	No	Yes	120
−52.0	*Pharus legumen*	λ	Yes	Yes	120
−52.2	*Pharus legumen*	λ	Yes	Yes	120
−54.0	*Ensiculus cultellus*	ν	Yes	Yes	120
−54.9	*Ensis directus*	α	No	Yes	120
−55.2	*Siliqua patula*	ρ	Yes	Yes	120
−55.2	*Siliqua patula*	π	Yes	Yes	120
−55.2	*Siliqua patula*	*o*	Yes	Yes	120
−55.2	*Ensis directus*	α	No	Yes	120
−55.2	*Ensis directus*	α	No	Yes	120
−56.6	*Ensis directus*	δ	No	Yes	120
−56.6	*Ensis minor* (Chenu)	η	Yes	Yes	120
−56.6	*Ensis macha*	θ	Yes	Yes	120
−56.6	*Ensis magnus*	η	Yes	Yes	120
−56.6	*Ensis directus*	γ	No	Yes	120
−56.7	*Ensis ensis*	η	Yes	Yes	120
ΔG (kcal mol^{-1}*)*	*Species*		*Linked to 5S rDNA?*	*Complete sequence?*	*m.s. (nts)*
U1 snRNA					
−70.1	*Siliqua patula*		Yes	Yes	165
−70.3	*Ensis minor* (Dall)		Yes	No	164
−83.9	*Aplysia californica*		No	Yes	161
−84.7	*Ensis macha* A		Yes	No	164
−85.4	*Ensis macha* B		Yes	Yes	164
−85.9	*Ensis minor* (Chenu)		Yes	Yes	164
−86.1	*Ensis directus*		Yes	No	164
−86.2	*Ensis siliqua*		Yes	Yes	164
−86.2	*Ensis magnus*		Yes	Yes	164
−86.2	*Ensis ensis*		Yes	Yes	164
−87.9	*Ensiculus cultellus*		Yes	Yes	164
−87.9	*Pharus legumen*		Yes	Yes	164
−91.3	*Lottia gigantea*		No	Yes	166

Abbreviations: IGS, intergenic spacer (downstream a U1 small nuclear RNA coding region); m.s., molecule size; NTS, non-transcribed spacer (intergenic spacer downstream a 5S ribosomal RNA coding region); nts, nucleotides; U1 snDNA, U1 small nuclear DNA; U1 snRNA, U1 small nuclear RNA; 5S rDNA, 5S ribosomal DNA; 5S rRNA, 5S ribosomal RNA.
Most positive ΔG values correspond to less stable structures. Bold values correspond to putative pseudogenised copies (see main text).

the supergroups with $n \geqslant 3$ (see Table 3). In supergroup I phylogeny (Figure 5a), each NTS group was recovered as monophyletic with high bootstrap support, and α- and β-sequences were included in a highly supported clade. In supergroup II phylogeny (Figure 5b), θ- and η-sequences (from mixed clones) were included in the same clade (bootstrap value of 100), with respect to a clade formed by γ- and δ-sequences. However, θ-sequences were very similar to γ-and δ-sequences in some regions of the alignment, and may represent an intermediate state between η- and γ-/δ-NTSs. The alignment of supergroup III sequences displayed an unexpected similarity between *E. macha* and *E. cultellus* NTSs (which appeared to be somewhat conserved among different genus). However, in the corresponding phylogeny, each NTS group was highly supported (Figure 5c). Supergroup IV alignment also revealed a certain degree of conservation among NTSs retrieved from different genus, and the phylogeny supported each NTS group with the highest value (Figure 5d).

Sequences considered for the U1 secondary structure prediction were subjected to phylogenetic analyses, excluding putative pseudogenised copies (Figure 6). *Ensis* U1 sequences were included in a nonsupported clade, and all of them, except the divergent *E. macha* type A U1 were recovered as monophyletic with a bootstrap support of 70. However, if the divergent sequence was excluded from the analysis (tree not shown), then the clade containing all remaining *Ensis* U1s decreased its bootstrap value. European *Ensis* sequences were grouped together with a bootstrap value of 91. Razor shell and gastropod sequences were reciprocally monophyletic with the highest support.

Two different phylogenies of IGS sequences were performed (Figure 7), one including supergroup *Ensis—Ensiculus* sequences, and another one including supergroup *Pharus—Siliqua* ones. The phylogeny of supergroup *Ensis—Ensiculus* (Figure 7a) recovered American and European *Ensis* sequences, as reciprocally monophyletic with high bootstrap support; the same happened with *Ensis* and *Ensiculus* sequences. In this tree, IGSs from the species *E. macha* were the ones located downstream type B U1s (no IGS downstream the type A U1 was sampled). The phylogeny of supergroup *Pharus—Siliqua* (Figure 7b) also recovered sequences from each species as monophyletic.

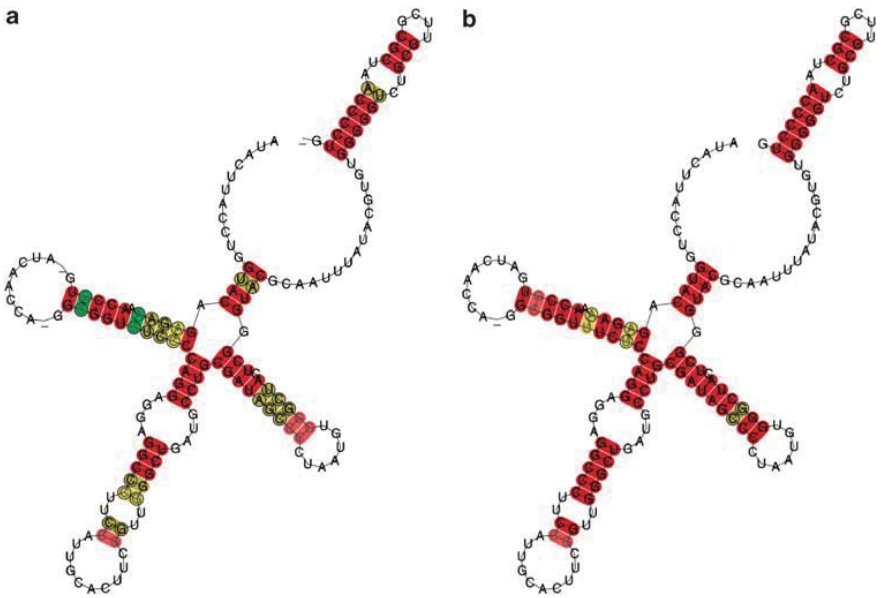


Figure 3 Predicted consensus secondary structure of U1 small nuclear RNA. Stem-loops are indicated by Roman numerals, following Lo and Mount (1990). Red, ochre and green indicates one, two and three types of base pairs, respectively. Pale colours indicate pairs that cannot be formed by all sequences. (**a**) Including razor shell and gastropod sequences. (**b**) Including only razor shell sequences.

Discussion

Long-term evolution of 5S rDNA in Pharidae

The 5S region of razor shells was not very polymorphic but the last three sites varied widely when considering the whole dataset. This means that the real number of variants in each species is higher than the number of predicted secondary structures obtained (because we only used complete 5Ss, after excluding the primer-annealing regions, for secondary structure prediction). We characterised several 5S sequences and found that a single species could have more than one 5S variant and that some of these variants were shared among species. Similarly, some NTSs were more closely related to NTSs from other species (and genera) than to NTSs from the species they were retrieved from. Therefore, the existence of divergent NTS sequences predates the speciation of the group. Several variants likely already occurred in the most recent common ancestor of the Pharidae (ancestral polymorphism) and some of them were retained until the present time.

The presence of pseudogenes in a multigene family strongly suggests that it evolves under a birth-and-death process (Rooney and Ward, 2005 and references therein). In this survey, we have found putative pseudogenised and truncated 5S copies. However, the long-term evolution of 5S rDNA in Pharidae appears to be a more complex issue. New variants arise through gene duplication, some of them are retained in the genome and others accumulate mutations and become pseudogenes (birth-and-death process). The action of purifying selection seems to be important to mantain the integrity of the RNA-coding regions and the upstream and downstream elements, and unequal crossing-overs and gene conversions should be also taking part and may be responsible for some of the sequence homogeneity (at least among 5S rDNA repeats located in the same array). We could suppose that divergent NTSs are located at different arrays where they have evolved independently, but we have shown that some species (*S. patula* and *E. cultellus*) have divergent NTS organised in tandem. In the same way, a few clones containing α- and δ-, and α- and γ- NTSs were characterised in *E. directus* (Vierna *et al*., 2009), and other studies found an intermixed organisation of 5S rDNA variants in grey mullets and *E. macha* (Gornung *et al*., 2007; Fernández-Tajes and Méndez, 2009). These findings support the idea that 5S rDNA frequently moves on within genomes (see below). If we consider that a 5S rDNA variant is the result of independent (nonconcerted) evolution in a given genomic location, this variant may have later been transposed into another array containing a different variant. Then, the intermixed organisation would be the result of duplications involving both variants, so they could eventually spread throughout the array. From our results, however, it is not clear whether divergent variants located in the same array are being more and more homogenised through the mechanisms typically involved in the concerted evolution of ribosomal multigene families. In conclusion, the long-term evolution of 5S rDNA in Pharidae has been mainly driven by birth-and-death processes and purifying selection. Nevertheless, homogenising mechanisms, such as unequal crossing-overs (favoured by the tandem organisation of 5S rDNA repeats) and gene conversions have been probably taking part, in agreement with what was previously reported on *Ensis* species (Vierna *et al*., 2009, 2010). Interestingly, recent studies in various animal groups have come to the same conclusion (Fujiwara *et al*., 2009; Freire *et al*., 2010; Úbeda-Manzanaro *et al*., 2010). Other techniques, such as fluorescent in situ hybridisa-

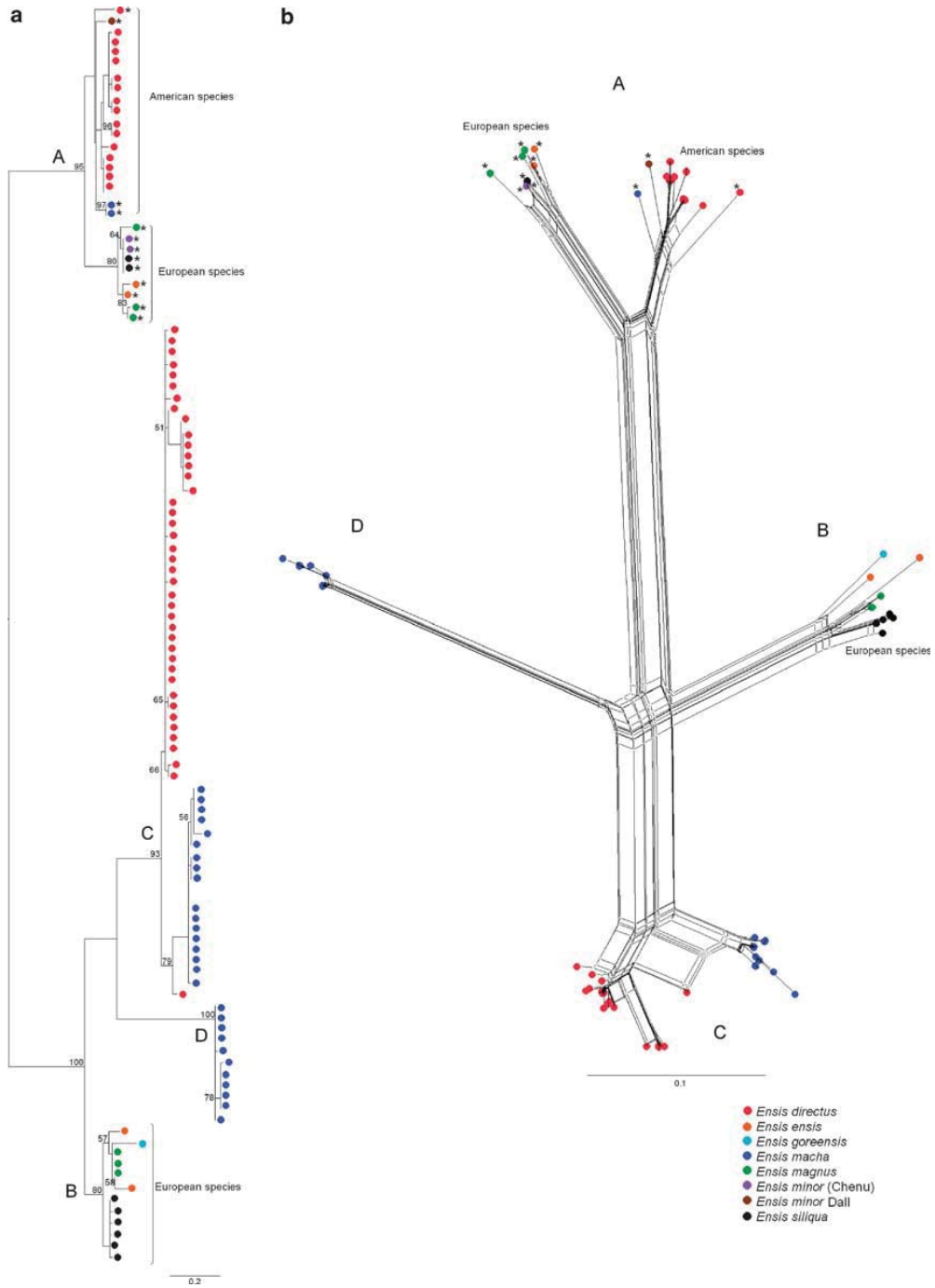


Figure 4 Phylogenetic relationships of the 100 nucleotides upstream the 5S ribosomal DNA transcription start site of *Ensis* species. Four different types of upstream regions (A–D) are identified. Upstream region (A) includes sequences from mixed clones of 5S ribosomal DNA and U1 small nuclear DNA of the European species, the American species and sequences from γ- and δ-NTSs from *E. directus*; (B) includes sequences from nonmixed clones of the European species; (C) sequences from the American species (*E. directus* α-NTSs and *E. macha* β-NTSs); and (D) sequences from *E. macha* ε-I and ε-II NTSs. The relationships among the different sequences are consistent with the phylogenetic history of the genus, as European and American species are reciprocally monophyletic (Vierna *et al*. unpublished). However, the phylogenetic pattern must be understood in the light of a birth-and-death evolutionary scenario (see main text). Asterisks (*) indicate the repeats retrieved from mixed clones. (**a**) Maximum likelihood phylogenetic tree constructed using the K80 + G model. Numbers on the tree correspond to nonparametric bootstrap supports (1000 replicates) and they are reported only for nodes with values ⩾50. Each upstream region type is indicated at the most external node common to all its sequences. (**b**) Phylogenetic network constructed using the neighbour-net algorithm and uncorrected P-distances. For NTS types, see Table 3.

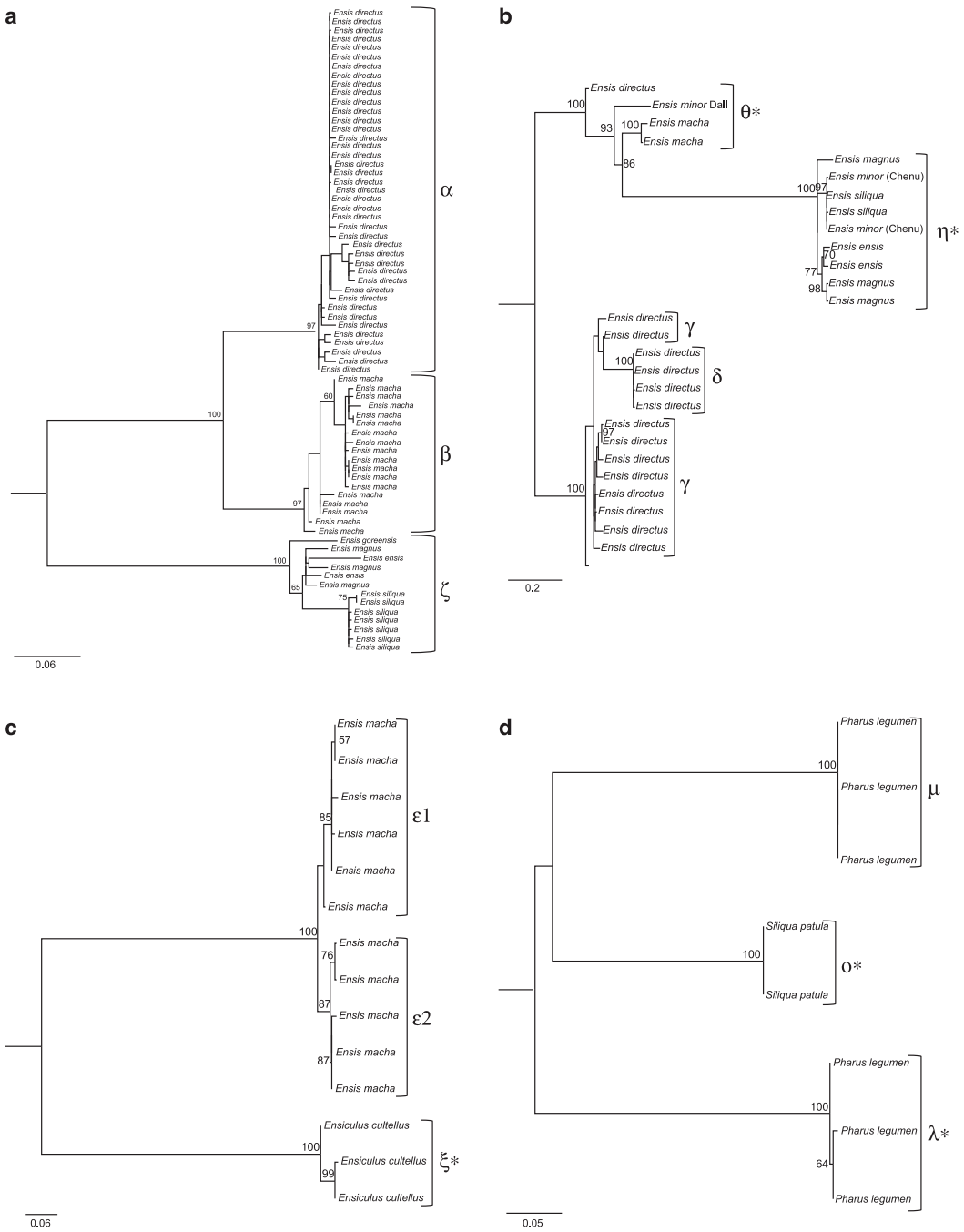


Figure 5 Maximum likelihood phylogenies of the nontranscribed spacers (NTSs) downstream the 5S ribosomal RNA coding regions of razor shell species. Numbers on the trees correspond to nonparametric bootstrap supports (1000 replicates) and they are reported only for nodes with values $\geqslant$50. NTSs groups (Table 3) are indicated. Asterisks (*) indicate NTS sequences retrieved from mixed clones of 5S ribosomal DNA and U1 small nuclear DNA. (**a**) Phylogeny of supergroup I NTSs reconstructed by the TVM + G model. (**b**) Phylogeny of supergroup II NTSs reconstructed by the TVM + G model. (**c**) Phylogeny of supergroup III NTSs reconstructed by the K81uf + G model. (**d**) Phylogeny of supergroup IV NTSs reconstructed by the GTR + G model.

tion may provide interesting data regarding the chromosomal locations of 5S rDNA arrays in razor shells and this should be an issue for further research.

U1 snDNA variation

We have characterised U1 snDNA for the first time in Bivalvia. Some of the species shared the same U1 variant,

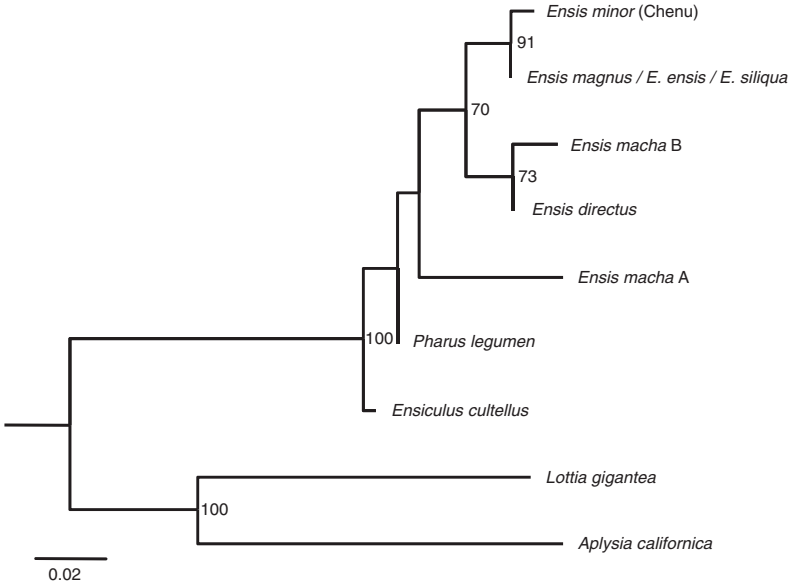


Figure 6 Maximum likelihood phylogeny of the U1 small nuclear RNA coding region of the razor shell and gastropod species, reconstructed using the K81 + I model. Sequences analysed correspond to putative functional copies on the basis of their predicted secondary structures and free energies. Numbers on the tree correspond to nonparametric bootstrap supports (1000 replicates). They are reported only for nodes with values $\geqslant 50$. *Ensis macha* type A and *E. directus* sequences were completed with the last 23 nts of the *E. magnus* sequence (see main text).

many others had a single U1 (not shared) variant and one of them (*E. macha*) had two different U1s, named type A and type B. The phylogeny of the U1 region places the *E. macha* type A sequence outside the clade formed by the other *Ensis* sequences. Taking into consideration that the relationships between this clade and the *E. cultellus*, *P. legumen* and *E. macha* type A sequences were not resolved, the latter one could be an old copy that diverged before the speciation of *Ensis*. However, it could well be a pseudogenised copy too (for example, derived from the type B U1) because the –25 region was different in two sites compared with the other *Ensis* sequences. We cannot be sure whether the predicted secondary structure was functional or not, as it was somewhat different compared with the other *Ensis* structures but had an intermediate ΔG value.

In the survey by Marz *et al*. (2008), discernible paralogs of spliceosomal snDNA multigene families were not uncommon within genera or families, but no dramatically different paralogs were found. We should take into account that we have only searched for tandemly repeated U1 snDNA (not found) or U1 snDNA linked to 5S rDNA. This means that dispersed U1 snDNA and U1 snDNA linked to other multigene families may occur in the genomes of Pharidae species, and these (hypothetical) copies and the ones linked to 5S rDNA would be paralogs.

We should be cautious regarding U1 snDNA long-term evolution because the number of repeats we obtained from each species was small. In any way, it is clear that duplication events and purifying selection have been involved.

Upstream elements, internal regulatory regions and downstream elements

The upstream elements, the internal regulatory regions and the termination signals are essential in 5S rDNA transcription, but epigenetic mechanisms were also found to be involved in transcription regulation (Douet and Tourmente, 2007). A TATA-like motif located at around –30 to −25 nt is essential for efficient transcription *in vitro* in *Caenorhabditis elegans* and *C. briggsae* (Nelson *et al*., 1998), *Neurospora crassa* (Tyler, 1987) and *D. melanogaster* (Sharp and Garcia, 1988). In razor shells, the TATA-like −25 region that we found upstream the 5S rDNA transcription start site is likely to be analogous to that of the mentioned organisms. Among the 5S internal regulatory regions, the ICR II was the most conserved one in the comparisons with *D. melanogaster* ICRs.

The transcription termination signal of 5S rDNA has been studied in various organisms and seems to be quite conserved (a TTTT stretch). We have found this element in almost all razor shell NTSs, in agreement with previous findings in other eukaryotes (Bogenhagen and Brown, 1981; Huang and Maraia, 2001).

According to Marz *et al*. (2008), the classical snDNA-specific proximal sequence elements (PSE) and TATA boxes that have been described in detail for several vertebrates and were highly conserved (Hernandez, 2001; Domitrovich and Kunkel, 2003) are the exception rather than the rule, as the snDNA promoters are highly diverse across metazoa. In *Drosophila*, there are two elements essential for the efficient initiation of transcription of snDNA families transcribed by RNA polymerase II: they are the PSEA (−61 to −41 nt), analogous to the vertebrate PSE and the PSEB (from −32 to −25 nt; consensus sequence C/TATGGAA/GA, Lo and Mount, 1990) (Zamrod *et al*., 1993). In razor shells and gastropods, we have identified an A/G-rich region (−25 region; from −27 to −23 nts), which was conserved in location and quite conserved in sequence that could correspond to the PSEB. The region centred at around −40 nts upstream the U1 snDNA transcription start site

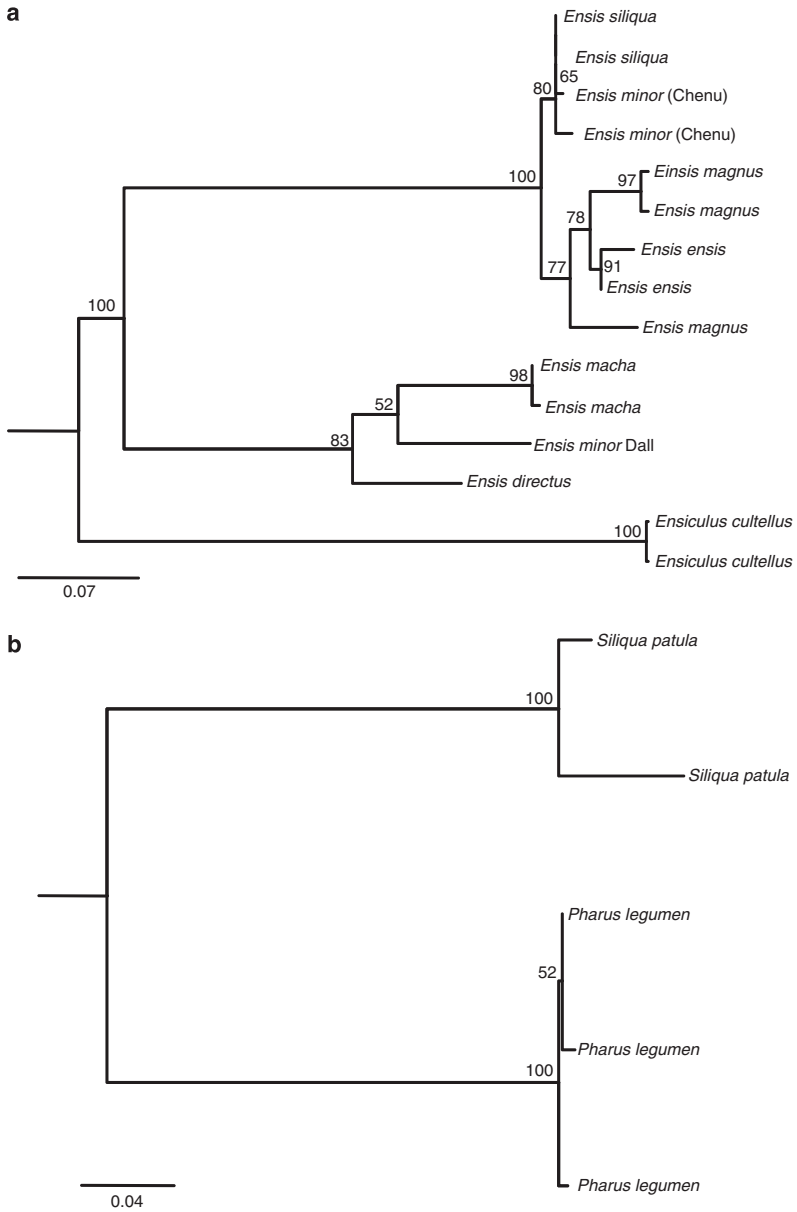


Figure 7 Maximum likelihood phylogenies of the intergenic spacers (IGS) downstream the U1 small nuclear RNA coding region. Numbers on the tree correspond to nonparametric bootstrap supports (1000 replicates). They are reported only for nodes with values $\geqslant 50$. (**a**) Phylogeny of supergroup *Ensis—Ensiculus* IGSs reconstructed following the HKY + G model. (**b**) Phylogeny of supergroup *Pharus—Siliqua* IGSs reconstructed following the HKY + G model. For IGS types, see Table 3.

was not very conserved, so it does not seem analogous to the PSEA/PSE. Interestingly, an AAAGC motif was found just upstream the U1 snDNA transcription start site of only some of our razor shell sequences (and not in the gastropod sequences). This pentanucleotide is shared with *D. melanogaster*, the slime mold *Physarum polycephalum* and some vertebrates (Lo and Mount, 1990 and reference therein). According to our data, it is not conserved in molluscs, but the occurrence of this motif in the same location and shared among distantly related taxa suggests it may have a function.

The internal regulatory regions within the U1 seemed to be somewhat more conserved than the ones within the 5S, as some of them were identical in the bivalve, gastropod, crustacean and insect species considered. Our data is consistent with the results by Marz *et al*. (2008), except in the positions 86, 131 and 132 (from the reference sequence, see Internal regulatory regions) that were not conserved in molluscs.

The snDNA transcription termination is more variable and different genes appear not to use a common process (Richard and Manley, 2009). For instance, transcription of

human U1 snDNA terminates close to the 3′ box (Cuello *et al*., 1999), but transcription of U2 snDNA extends about 800 sites beyond it (Medlin *et al*., 2003). The human 3′ box (Hernandez, 1985) is a 16 nt stretch, located 10 sites downstream the U1. In razor shells, we have not found a conserved region analogous to the 3′ box; however, the first five nts of the IGS were identical in *Ensis* species and *E. cultellus* and similar in *P. legumen*. Sequences from *S. patula* were somewhat different, but this could be related to the fact that their preceeding U1s were likely to be pseudogenised copies.

One or more linkage events throughout evolution?

In order to study whether the linkage happened once or more throughout the evolution of the Pharidae lineages, we have constructed several phylogenies and carefully studied the alignments performed. By mapping the 5Ss from mixed clones on the phylogenetic trees and on the network performed (Supplementary file S7), we tried to detect whether the linkage between the multigene families emerged once or more throughout the evolution of razor shells, but unfortunately, the phylogenies were not resolved.

The alignment of the IGS region supports that the linkage between both multigene families is homologous in these Pharidae species, with the exception of *E. macha* type A U1. In this case, as we did not sample its downstream IGS, we do not know how similar it would be compared with the other *Ensis* IGSs. The origin of this clone is unclear, as it could represent a new linkage between both multigene families, or it could be a descendant of the original linkage in which the NTS was replaced.

The most parsimonious explanation for our data is an evolutionary scenario in which the linkage happened only once, in a common ancestor to all the Pharidae species studied. Subsequently, there were duplications involving either the entire linked unit, or any of the RNA coding regions explaining why we have found the different genomic organisations recorded in Figure 1. Sequences started to accumulate mutations and diverged, but purifying selection and, perhaps, other homogenising mechanisms, maintained the integrity of the functional regions. Finally, different units continued to be duplicated and/or deleted across the different Pharidae lineages.

How 5S rDNA and U1 snDNA can become linked and why?

There are two possible alternatives for both multigene families to become linked: the linkage was the consequence of the insertion of one or more 5S rDNA repeats next to one or more U1 snDNA repeats, or vice-versa. However, how could this happen? Several surveys have suggested that rDNA (both 5S rDNA and the major ribosomal genes) frequently moves on from one location to another in the eukaryote genome (Rooney and Ward, 2005; Datson and Murray, 2006; Veltos *et al*., 2009; Nguyen *et al*., 2010), and several mechanisms have been proposed to explain this apparent mobility. Drouin and Moniz de Sá (1995) hypothesised that a 5S rDNA transposition could be produced at the DNA level mediated by extrachromosomal circular DNA or by an RNA intermediate. Interestingly, recent surveys have given support to both hypothesis (Kalendar *et al*., 2008; Cohen *et al*., 2010). Similarily, Rooney and Ward, (2005) hypothesised that 5S rDNA was capable of multiplying and integrating into other areas of the genome through a process the same as, or similar to, retroposition, in filamentous fungi. In a survey concerning the major ribosomal genes, it has been proposed that ectopic recombination (homologous recombination between repetitive sequences of nonhomologous chromosomes) was the primary motive force in the repatterning of these genes in lepidopteran species (Nguyen *et al*., 2010). Similar to what has been reported for rDNA, Marz *et al*., (2008) concluded that metazoan spliceosomal snDNA families behave like mobile genetic elements because they barely appear in syntenic positions, as measured by their flanking regions. Therefore, in theory, there are a few possible ways by which 5S rDNA and U1 snDNA could have become linked, but why?

Several examples have been reported in which 5S rDNA and U1 snDNA were found linked to each other or to other multigene families in virtually all eukaryote groups. Interestingly, the linkage between 5S rDNA and other multigene families have been repeatedly established and lost throughout evolution in several lineages, but this lack of conservation and the diversity of the linkages make it unlikely that they provide any selective advantage (for example, transcriptional co-regulation, (Drouin and Moniz de Sá, 1995). In the same way, Marz *et al*., (2008) concluded that tandem repeats of different spliceosomal snDNA families, or of a spliceosomal snDNA family and 5S rDNA, are not conserved over long evolutionary timescales in metazoans. So, even though the linkages between multigene families may provide a benefit that has not been reported yet, they rather seem to us to be the result of stochastic processes within genomes. The high copy number of 5S rDNA would make it quite likely to establish a linkage with another multigene family.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Jane Frydenberg and Camilla Håkansson, who were very kind and helpful during lab work in Aarhus Manja Marz for providing gastropod U1 snDNA sequences, her support during bioinformatic analyses and her comments on the paper; Rudo von Cosel for his help regarding razor shell taxonomy and the identification of many of the specimens studied in this survey; Miguel Vizoso for his comments on the manuscript; and Klaus Brehm for providing information about the *Taenia* spliced leader sequence. This work would have not been possible without the help of the following colleagues (alphabetical order) who kindly provided razor shell specimens: Dan Ayres, Emile Égea, John Havenhand, Iben Heiner, Inés Naya, Lobo Orensanz, Roberto Portela-Míguez, Anja Schulze, John Taylor, Ray Thompson, and Katrine Worsaae. JV has been supported by a 'María Barbeito' fellowship and a travel grant, both from the *Consellería de Economía e Industria*, *Xunta de Galicia* (Spain) and the European Social Fund.

References

Bailey TL, Elkan C (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Second Int Conf Intell Syst Mol Biol* pp. 28–36. AAAI Press, Menlo Park, California.

Barciszewska MZ, Szymanski M, Erdmann VA, Barciszewski J (2000). 5S Ribosomal RNA. *Biomacromolecules* **1**: 297–302.

Barzotti R, Pelliccia F, Rocchi A (2003). Identification and characterization of U1 small nuclear RNA genes from two crustacean isopod species. *Chromosome Res* **11**: 365–373.

Bogenhagen DF, Brown DD (1981). Nucleotide sequences in *Xenopus* 5S DNA required for transcription termination. *Cell* **24**: 261–270.

Branlant C, Krol A, Ebel JP, Lazar E, Haendler B, Jacob M (1982). U2 RNA shares a structural domain with U1, U4, and U5 RNAs. *EMBO J* **1**: 1259–1265.

Brehm K, Hubert K, Sciutto E, Garate T, Frosch M (2002). Characterization of a spliced leader gene and of trans-spliced mRNAs from *Taenia solium*. *Mol Biochem Parasitol* **122**: 105–110.

Bryant D, Moulton V (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**: 255–265.

Cabral-de-Mello DC, Moura RC, Martins C (2010). Chromosomal mapping of repetitive DNAs in the beetle *Dichotomius geminatus* provides the first evidence for an association of 5S rRNA and histone H3 genes in insects, and repetitive DNA similarity between the B chromosome and A complement. *Heredity* **104**: 393–400.

Caradonna F, Bellavia D, Clemente AM, Sisino G, Barbieri R (2007). Chromosomal localization and molecular characterization of three different 5S ribosomal DNA clusters in the sea urchin *Paracentrotus lividus*. *Genome* **50**: 867–870.

Chen S-H, Su S-Y, Lo C-Z, Chen K-H, Huang T-J, Kuo B-H *et al.* (2009). PALM: a paralleled and integrated framework for phylogenetic inference with automatic likelihood model selectors. *PLoS ONE* **4**: e8116.

Cohen S, Agmon N, Sobol O, Segal D (2010). Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. *Mobile DNA* **1**: 11.

Cosel von R (1993). The razor shells of the eastern Atlantic, part 1: Solenidae and Pharidae I. *Arch Moll* **122**: 207–321.

Cosel von R (2009). The razor shells of the eastern Atlantic, part 2. Pharidae II: the genus *Ensis* Schumacher, 1817 (Bivalvia, Solenoidea). *Basteria* **73**: 1–48.

Cross I, Rebordinos L (2005). 5S rDNA and U2 snRNA are linked in the genome of Crassostrea angulata and Crassostrea gigas oysters: does the (CT)n·(GA)n microsatellite stabilize this novel linkage of large tandem arrays? *Genome* **48**: 1116–1119.

Cuello P, Boyd DC, Dye MJ, Proudfoot NJ, Murphy S (1999). Transcription of the human U2 snRNA genes continues beyond the 39 box *in vivo*. *EMBO J* **18**: 2867–2877.

Datson PM, Murray BG (2006). Ribosomal DNA locus evolution in *Nemesia*: transposition rather than structural rearrangement as the key mechanism? *Chrom Res* **14**: 845–857.

Domitrovich AM, Kunkel GR (2003). Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. *Nucleic Acids Res* **31**: 2344–2352.

Douet J, Tourmente S (2007). Transcription of the 5S rRNA heterochromatic genes is epigenetically controlled in *Arabidopsis thaliana* and *Xenopus laevis*. *Heredity* **99**: 5–13.

Drouin G, Moniz de Sá M (1995). The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol Biol Evol* **12**: 481–893.

Eickbush TH, Eickbush DG (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**: 477–485.

Freire R, Arias A, Insua AM, Méndez J, Eirín-López JM (2010). Evolutionary dynamics of the 5S rDNA gene family in the mussel *Mytilus*: mixed effects of birth-and-death and concerted evolution. *J Mol Evol* **70**: 413–426.

Felsestein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Fernandez M, Ruiz ML, Linares C, Fominaya A, de la Vega MP (2005). 5S rDNA genome regions of *Lens* species. *Genome* **48**: 937–942.

Fernández-Tajes J, Méndez J (2009). Two different size classes of 5S rDNA units coexisting in the same tandem array in the razor clam *Ensis macha*: is this region suitable for phylogeographic studies? *Biochem Genet* **47**: 775–788.

Fujiwara M, Inafuku J, Takeda A, Watanabe A, Fujiwara A, Kohno S *et al.* (2009). Molecular organization of 5S rDNA in bitterlings (Cyprinidae). *Genetica* **135**: 355–365.

Gascuel O (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14**: 685–695.

Giegerich R, Meyer F, Schleiermacher C (1996). GeneFisher—software support for the detection of postulated genes. *Proc Int Conf Intell Syst Mol Biol* **4**: 68–77.

Gornung E, Colangelo P, Annesi F (2007). 5S ribosomal RNA genes in six species of Mediterranean grey mullets: genomic organization and phylogenetic inference. *Genome* **50**: 787–795.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**: 95–98.

Hernandez N (1985). Formation of the 3′ end of U1 snRNA is directed by a conserved sequence located downstream of the coding region. *EMBO J* **4**: 1827–1837.

Hernandez N (2001). Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J Biol Chem* **276**: 26733–26736.

Hofacker IL (2003). Vienna RNA secondary structure server. *Nucleic Acids Res* **31**: 3429–3431.

Huang Y, Maraia RJ (2001). Comparison of the RNA polymerase III transcription machinery in *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae* and human. *Nucleic Acids Res* **29**: 2675–2690.

Huson DH, Bryant D (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**: 254–267.

Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O *et al.* (2008). *Cassandra* retrotransposons carry independently transcribed 5S RNA. *PNAS* **105**: 5833–5838.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Little RD, Braaten DC (1989). Genomic organization of human 5 S rDNA and sequence of one tandem repeat. *Genomics* **4**: 376–383.

Lo PCH, Mount SM (1990). *Drosophila melanogaster* genes for U1 snRNA variants and their expression during development. *Nucleic Acids Res* **18**: 6971–6979.

Manchado M, Zuasti E, Cross I, Merlo A, Infante C, Rebordinos L (2006). Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: a new linkage to the U1, U2, and U5 small nuclear RNA genes. *Genome* **49**: 79–86.

Marz M, Kirsten T, Stadler PF (2008). Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol* **67**: 594–607.

Mathews DH, Sabina J, Michael Zuker M (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911–940.

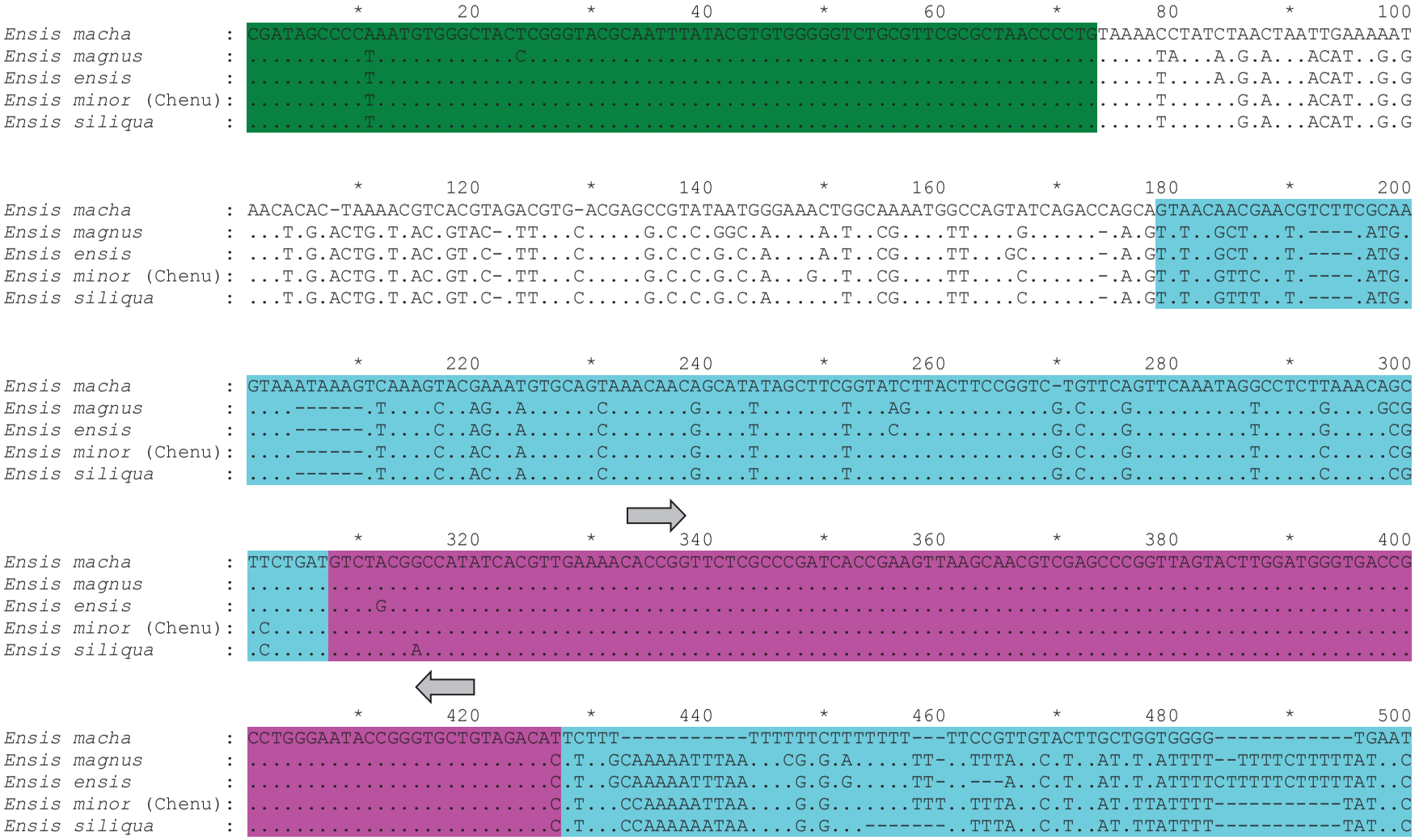
Medlin JE, Uguen P, Taylor A, Bentley DL, Murphy S (2003). The C-terminal domain of pol II and a DRB-sensitive kinase are required for 3′ processing of U2 snRNA. *EMBO J* **22**: 925–934.

Morzycka-Wroblewska E, Selker EU, Stevens JN, Metzenberg RL (1985). Concerted evolution of dispersed *Neurospora crassa*

5S RNA genes: pattern of sequence conservation between allelic and nonallelic genes. *Mol Cell Biol* **5**: 46–51.
Mount SM, Gotea V, Lin C-F, Hernandez K, Makałowski W (2007). Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA* **13**: 5–14.
Nei M, Rooney AP (2005). Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**: 121–152.
Nelson DW, Linning RM, Davison PJ, Honda BM (1998). 5′-flanking sequences required for efficient transcription *in vitro* of 5S RNA genes, in the related nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Gene* **218**: 9–16.
Nguyen P, Sahara K, Yoshido A, Marec F (2010). Evolutionary dynamics of rDNA clusters on chromosomes of moths and butterflies (Lepidoptera). *Genetica* **138**: 343–354.
Pelliccia F, Barzotti R, Bucciarelli E, Rocchi A (2001). 5S ribosomal and U1 small nuclear RNA genes: a new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences. *Genome* **44**: 331–335.
Peterson RC, Doering JL, Brown DD (1980). Characterization of two *Xenopus* somatic 5S-DNAs and one minor oocyte-specific 5S-DNA. *Cell* **20**: 131–141.
Pieler T, Hamm J, Roeder RG (1987). The 5S gene internal control region is composed of three distinct sequence elements, organized as two functional domains with variable spacing. *Cell* **48**: 91–100.
Posada D, Crandall KA (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.
Query CC, Bentley RC, Keene JD (1989). A specific 31-nucleotide domain of U1 RNA directly interacts with the 70K small nuclear ribonucleoprotein component. *Mol Cell Biol* **9**: 4872–4881.
Reuter JS, Mathews DH (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**: 129.
Richard P, Manley JL (2009). Transcription termination by nuclear RNA polymerases. *Genes Dev* **23**: 1247–1269.
Rooney AP, Ward TJ (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. *PNAS* **102**: 5084–5089.
Scherly D, Boelens W, van Venrooij WJ, Dathan NA, Hamm J, Mattaj IW (1989). Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. *EMBO J* **8**: 4163–4170.
Sharp S, Garcia A, Cooley L, Söll D (1984). Transcriptionally active and inactive gene repeats within the *D. melanogaster* 5S RNA gene cluster. *Nucleic Acids Res* **20**: 7617–7632.
Sharp SJ, Garcia AD (1988). Transcription of the *Drosophila melanogaster* 5S RNA gene requires an upstream promoter and four intragenic sequence elements. *Mol Cell Biol* **8**: 1266–1274.
Shippen-Lentz DE, Vezza AC (1988). The three 5S rRNA genes from the human malaria parasite *Plasmodium falciparum* are linked. *Mol Biochem Parasit* **27**: 263–273.
Swofford DL (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version*. Sinauer Associates, Sunderland, MA, USA.
Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular Evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.
Tyler BM (1987). Transcription of *Neurospora crassa* 5 S rRNA genes requires a TATA box and three internal elements. *J Mol Biol* **196**: 801–811.
Úbeda-Manzanaro M, Merlo MA, Palazón JL, Sarasquete C, Rebordinos L (2010). Sequence characterization and phylogenetic analysis of the 5S ribosomal DNA in species of the family Batrachoididae. *Genome* **53**: 723–730.
Veltos P, Keller I, Nichols RA (2009). Geographically localised bursts of ribosomal DNA mobility in the grasshopper *Podisma pedestris*. *Heredity* **103**: 54–61.
Vierna J, González-Tizón AM, Martínez-Lage A (2009). Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). *Biochem Genet* **47**: 635–644.
Vierna J, Martínez-Lage A, González-Tizón AM (2010). Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. *Genome* **53**: 23–34.
Will CL, Lührmann R (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem* **386**: 713–724.
Zamrod Z, Tyree CM, Song Y, Stumph WE (1993). *In vitro* transcription of a *Drosophila* U1 small nuclear RNA gene requires TATA box-binding protein and two proximal cis-acting elements with stringent spacing requirements. *Mol Cell Biol* **13**: 5918–5927.
Zhuang Y, Weiner AM (1986). A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell* **46**: 827–835.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)

Supplementary File S1 DNA alignment of the U1-U1 clones from *Ensis* species. Regions shaded in green are partial RNA coding regions of U1 small nuclear DNA. Light blue and grey shaded regions are similar to delta and gamma nontranscribed spacers from the species *E. directus* (see main text). The fuchsia region corresponds to the RNA coding region of 5S ribosomal DNA (5S). Grey region is possibly an old pseusogenised 5S copy, as it is similar to the 5S region flanked by the arrows. Dots represent residues identical to those of the first sequence, dashes indicate alignment gaps, and base substitutions are indicated by the respective base.



```
                          *       520         *       540         *       560         *       580         *       600
Ensis macha         : ATTCATTTTACTGCCTTCTAATACAATTAATCTATTTGTCTGGTGTCTAGACAAGTCTCGT-----GGGTGTTATG------TTCGCAGCACCGTTAATC
Ensis magnus        : ....CC.GC...TG..GG.GGGGT..AC.T..GC..GACGGTT...T.TT.T.CT.TCG.CTATTAAAC....G.CTGACGC...A.....T.G..C..
Ensis ensis         : ....CC.GC...TG..GG.GGGGT..AG.T.TGC..GACGGTT...T.TT.T.CT.TCG.CTATTAAAC....G.CTGACGC...A.....TAG..C..
Ensis minor (Chenu): ....CC.GC...TG..GGGGGGGT..AG.T..GC...ACAATC...T.TT.T.CT.TCG.CTATTAAAC....G.CCGACGC...A.....TAG..C..
Ensis siliqua       : ....CC.GC...TG..GG.GGGGT..AG.T..GC...ACGATC...T.TT.T.CT.TCG.CTATCAAAC....G.CCGACGC...A.....TAG..C..

                          *       620         *       640         *       660         *       680         *       700
Ensis macha         : ACCTTAGTATAGAGGTTAAGCAACGCCTTGTGC-CGTCAGTTCTTGGATGGGTTTCC-TTCGAATCGGCCGGGTAATG-TATAACATTACATTGTCTGCT
Ensis magnus        : .A.G..T..CT..C.........T.TT....C.CG......A..........GA-..G..T.CGAT........CC..A.G.C.A...C.-..T.T.--.
Ensis ensis         : .A.G..T..CT..C..........TT....C.-G......A..........GAA..G..T.CGAT........CC..A.G.C.A...C.G..T.TC--C
Ensis minor (Chenu): .A.G..T..CT..A...........T.....C.-G......A..........GGA..G..T.CGAT........CC..A.G...A...C.G..T.T.T.C
Ensis siliqua       : .A.G..T..CT..A...........T.....C.CG......A..........GGA..G..T.CGAT........C-..A.G...A---T.G..T.T.--C

                          *       720         *       740         *       760         *       780         *       800
Ensis macha         : TTTTTTTTCTTT--TCAAGTTGTAAATGAGTTTCGTTCTCGTTGTGCTTTGT--TTGCGTGCAGCACCAGGAACAAAGGGAATACAAACAAAAGTACGTA
Ensis magnus        : C..CA.G.T-..--.TGG.G...GCG..T.AC--A-.G...GCACA.AC--ACAGCA.ACA......ACAC.CAC.C.AC.----C.C------------
Ensis ensis         : CC.CA.G.TC..--.TTG.G...GCG..T.AAG-A-.G...G..AA.AACAACAACAACAA..A..A..AC.....CAAC..C.AC.C..GC..C.ACA.
Ensis minor (Chenu): CC.CA.GCT...TA.TTG.G...GCG..T.AAG-AA.G...GGTAA.AACCA-ACAA.CAA..A..A..AC.C...CAAC.CC..C.....C.CC..CA.
Ensis siliqua       : CC.CA.GCT...-A.TTG.G...GCG..T.AAG-A-.G...G.ACA.AAC------A.CAA..A..A..AC.C...CAAC.CC..C.....C..C..CA.

                          *       820         *       840         *       860         *       880         *       900
Ensis macha         : CAACAATGTTTTTCGCGAATAAAAATTATACACATGCACTGAAATATTGATTCGTTGAATATTGGTTTCGATGATTGCTTCTGGTGATGGATTGCTTGAT
Ensis magnus        : --...CGAC-----A.-.GC..C.GCA.C.ACA.C.G.A...T..T...G.G.A....--------..TA.A...G.AA.TGAAA....AT...AA-..C
Ensis ensis         : .....CCACCACCAA.-..C..C.GCA.C..CA.C.G.A...T..G.....G.A....--------..TA.A.....AA.TGAAA....AT...AA-..C
Ensis minor (Chenu): ......CA.------.-..C.GC..CA.C.ACA.C.G.A...T..T.....G.A....--------..TA.A...G.AA.TGAAA....AC...AA-..C
Ensis siliqua       : ......CA.------.-..C.GC..CA.C.ACA.C.G.A...T..T.....G.A....--------..GA.A...G.AA.TGAAA....AC...AA-..C

                          *       920         *       940         *       960         *       980         *      1000
Ensis macha         : CGATTGATTGAGT-TGATTGATTGATTGAT-TGATTGATTGATTGACTGACTTCCCTTCTCTTATCTTCATACAACAACATACATACATACATACATACA
Ensis magnus        : T....C.A.ACA.G.A.A....GA.....CG...CGTG........TGA.T.GAT-----GAA..GA.TGATTG.TG------------------------
Ensis ensis         : T....C.A.A.A.G.A.A....GA.....CG...CGTG........T.A.T.GAT-----.AA..GA.TGATTG.TG------------------------
Ensis minor (Chenu): T....C.A.A.A.GGA.A....GA.....CG...CGTG........TGA.T.GATT.GA.GAA..GA.TGATTG.TTG-..TG..TG..TG..TG..TG.
Ensis siliqua       : T....C.A.A.A.GGA.A....GA.....CG...CGTG........TGA.T.GATT.GA.GAA..GA.TGATTG.TTG-..TG..TG.--------.TG.

                          *      1020         *      1040         *      1060         *      1080         *      1100
Ensis macha         : TACATACATACATACATACATACATACATACATACATTTTTCTTTTCTTTTCTTTCTACAGGGCTTCGAAGCAAAGACGTGTGACCCTGGCGACGTTCCC
Ensis magnus        : --T..TA..GGC.G---G...TAT.TTT....GGGCGGGACAAAGAACACGGCCA.G...A.CAC.G.G.A...TTCGACT.T.---..A.ATG.....-
Ensis ensis         : --T..TA..GGC.G---G...TAT.TTT....GGGCGGGACAAAGAACACGGCCA.G...A.CAC.G.G.A...TTCGACT.T.---..A.ATG.....-
Ensis minor (Chenu): .GT..CA..GGC.G---G.C.TA..TTT....GGGCGGGACAAAGAACACGGCCA.G...A.CAC.A.G.A.G.GTCGACT.T.---.TA.ATG.....-
Ensis siliqua       : .GT..CA..GGC.G---G.C.TA..TTT....GGGCGGGACAAAGAACACGGCCA.G...A.CAC.A.G.A.G.GTCGACT.T.---.TA.ATG.....-
```

```
                                *      1120         *      1140         *      1160         *      1180         *      1200
Ensis macha         : ATGCTTGAGTAGGTACGTCATCGTTTTAACATTCCTCCACTTTTCTCCTTATTTGCCGGCCTTCACATTCTGCTTACGGGTAGTCTATTTAGAACGTTGT
Ensis magnus        : -..A.G...---...T..AG........TT..GT....GT.G.AT.GA.A.G..TTGC.T...T.....A..T..C..T..----CCA.G.A........
Ensis ensis         : -..A.G...---...T..AG.........T..GTA...GT.G.AT.GA.A.G..TTGC.T...T.T...A..T..C..T..----CCA.G.A........
Ensis minor (Chenu): -.AA.C...---...T..AG............G.....GT.G.AT.GA.A.G..TT.C.T...T.....A..T..C..T..----CC.AG.A........
Ensis siliqua       : -.AA.C...---...T..AG.........T..G.....GT.G.AT.GA.A.G..TT.C.T...T.....A..T..C..T..----CC.AG.A........

                                *      1220         *      1240         *      1260         *      1280         *      1300
Ensis macha         : GAAACGTTTGTACCAGACAC--------------AAGTTGTGTCCGAAGCAGTTGACTAGTCCTTGTCAGATACTGCTCTCGTGGTAAAAAACGATAGAT
Ensis magnus        : .......A.T....CA....CGCACACCAACGTC.C......C..........A..T.T...T.G..A.T...TC..A.C.......C.G...T.A....
Ensis ensis         : .......AAT....CA....CGCACACCAACGTC.C......C..........A..T.C...T.G..A.T...TC..A.C.......C.G...T.A....
Ensis minor (Chenu): .......A.T....CA....CGCACACCAACGTC.C......C.....A....A..T.TA..T.G..A.T...TC..A.C.......C.G...T.A....
Ensis siliqua       : .......A.T....CA....CGCACACCAACGTC.C......C.....A....A..T.TA..T.G..A.T...TC..A.C.......C.G...T.A....

                                *      1320         *      1340         *      1360         *      1380         *      1400
Ensis macha         : TGTATCACTAAATTGAAAAATCAGTGGTGTGGAGTCAATTCGAAAA-TCACTGTCCGACACGTCGGGCCGAATAATACCCTTGTGTGAAAAGGTATGCAC
Ensis magnus        : AAA.A-...T.....CC.TT....A............T.CGA....AA..A..CA.A.....C....T.......CGA.AGCT.C.C......GT...G.
Ensis ensis         : AAA.A--..T.....CC.TT....A............T.CGA....AC..A..CA.A.....C....T.......CGA.AGCT.C.C......GT...G.
Ensis minor (Chenu): AAA.A-...T.....CC.TT....AT..C........T.CGA....ACA.A..CA.A.....C....TT......CGA.AGCT.C.C......GT...G.
Ensis siliqua       : AAA.A-...T.....CC.TT....AT..C........T.CGA....ACA.A..CA.A.....C....T.......CGA.AGCT.C.C......GT...G.

                                  *      1420         *      1440         *      1460
Ensis macha         : GTGCATAGGAAGGCATACTTACCTGGTACAGAGAATACCGTGATCAATCAGGCGGTTTCTCC
Ensis magnus        : A.....G.C..A..........................GA..........CC..............
Ensis ensis         : A...C.G.C..A..........................GA..........CC..............
Ensis minor (Chenu): C.....G.C..A.................G....GA..........CC..............
Ensis siliqua       : C.....G.C..A..........................GA..........CC..............
```

Supplementary File S2 DNA stretches (not aligned) of 100 nucleotides upstream the transcription start sites of 5S ribosomal DNA (5S rDNA) and U1 small nuclear DNA (U1 snDNA). A conserved region (-25 region) was identified, located at positions -28 to -22 nucleotides. Shaded nucleotides correspond to identical nucleotides respect to the first sequences. Yellow motifs are upstream the 5S rDNA transcription start site, and light blue ones are upstream the U1 snDNA transcription start site. The motifs AAAGC found just upstream the U1 snDNA transcription start site (and shared with *Drosophila melanogaster* and other organisms, see Discussion) are underlined. Fuchsia nucleotides constitute a less conserved region upstream the U1 snDNA transcription start site. *Ensis directus* sequences 1-41 belong to alpha nontranscribed spacer (NTS) sequences, and 42-56, to delta and gamma NTSs (see main text). (#) indicates sequence upstream a putative pseudogenised copy; (*) sequence retrieved from a mixed clone of 5S rDNA and U1 snDNA; (→) sequence upstream the *E. macha* type A U1 snDNA (see main text).

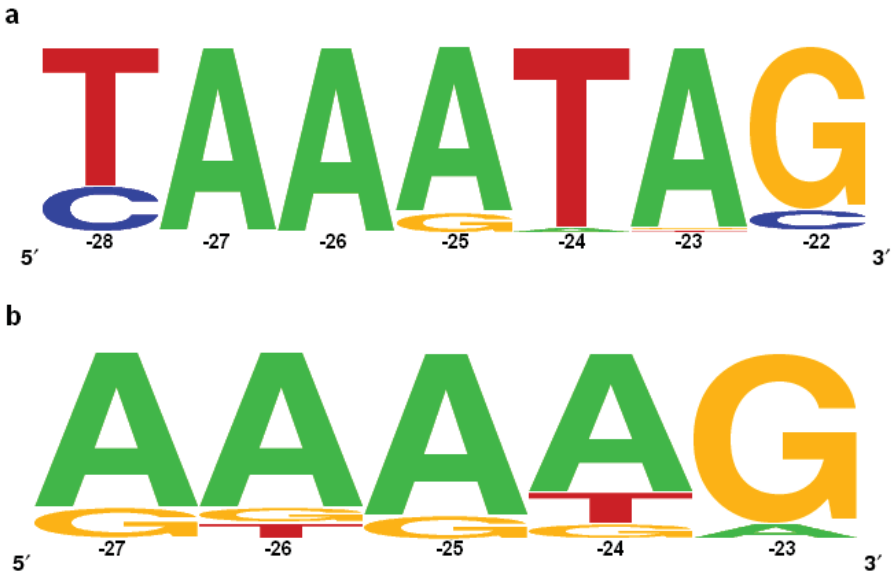
```
                                  *        -80        *        -60        *        -40        *        -20        *       -1
#Ensis directus 1       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 2       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 3       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 4       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCCTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 5       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCCTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 6       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 7       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 8       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 9       : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 10      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 11      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 12      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 13      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 14      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 15      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 16      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATATGTTAGTGTACCTCTT
 Ensis directus 17      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTCTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 18      : CGTGTCTACTTCACTCACCCGACACATGTATACATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTGGTGTCCCTCTT
 Ensis directus 19      : CTTCATGACATTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 20      : CTTCATGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 21      : TCTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 22      : CTACTTCGTGACCTTGACTCACCTGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 23      : ACTTCATGACCTTGACTCACCCGACACATGTATATGTTCTTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 24      : ACTTCATGACCTTGACTCACCCGACACATGTATATGTTCTTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 25      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 26      : ACTTCATGACCTTGACTCACCCGACACATGTATATGTTCTTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 27      : ACTTCATGACCTTGACTCACCCGACACATGTATATGTTTGTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 28      : ACTTCATGACCTTGACTCACCCGACACATGTATATGTTCTTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 29      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATACGCATGTGTTAGTGTACCTCTT
 Ensis directus 30      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 31      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 32      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 33      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 34      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 35      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 36      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 37      : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
```

```
 Ensis directus 38     : CTACTTCGTGACCTTGACTCACCCGACACATGTATGTTCTTTCTGAGAAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 39     : ACTTCATGACCTTGACTCACCCGACACATGTATATGTTCTTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 40     : ACTTCGTGACCTTGACTCACCCGACACATGTATATGTTCTTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 41     : CTACTTCGTGACCTTGACTTACCCGACACATGTATGTTCTTTGTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTACCTCTT
 Ensis directus 42     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGTACTCTTTAACACCTTCTGAT
 Ensis directus 43     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGTACTCTTTAACACCTTCTGAT
 Ensis directus 44     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGTCCTCTTTAACACCTTCTGAT
 Ensis directus 45     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGTCCTCTTTAACACCTTCTGGT
 Ensis directus 46     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGCCCTCTTTAACACCTTCTGAT
 Ensis directus 47     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGTCCTCTTTAACACCTTCTGAT
 Ensis directus 48     : GAAATTGTAGTACGAAATGTGCAGTAAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGCCCTCTTAAACAGCTTCTGAT
 Ensis directus 49     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGTCCTCTTTAACACCTTCTGAT
 Ensis directus 50     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGCCCTCTTTAACACCTTCTGAT
 Ensis directus 51     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGTCTGTTCGGTTCAAATAGTCCTCTTAAGTAGCTTCTGAT
 Ensis directus 52     : GAAATTGTAGTACGAAATGTGCAGTAAACAACAGCATATAGCTTCGGTATCTTACTTCCGGACTGTTCAGTTCAAATAGCCCTCTTAAACAGCTTCTGAT
 Ensis directus 53     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGTCTGTTCAGTTCAAATAGTCCTCTTAAACGGCTTCTGAT
 Ensis directus 54     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGTCTGTTCAGTTCAAATAGTCCTCTTAAACGGCTTCTGAT
 Ensis directus 55     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGTCTGTTCAGTTCAAATAGTCCTCTTAAACGGCTTCTGAT
 Ensis directus 56     : GAAATTGAAGTACGAAATGTGCAGTGAACAACAGCATATAGCTTCGGTATCTTACTTCCGGTCTGTTCAGTTCAAATAGTCCTCTTAAACGGCTTCTGAT
*Ensis directus 57     : AAGATTGAAGTACGAAATGTGCAGTAAACAACAGCATATAGCTTCGGTATCGTACTTCCGGTCTGTTCAGTTCAAATAATCCTCCTTAGCACCGTCTGAT
 Ensis macha 1         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATATGTTAGTGTCCCTCGT
 Ensis macha 2         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATATGTTAGTGTCCCTCGT
 Ensis macha 3         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATATGTTAGTGTCCCTCGT
 Ensis macha 4         : GTATTGTGTTAACTTCACTCACCCCGACACTGTATATTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 5         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 6         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 7         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 8         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 9         : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 10        : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTATCCCTCGT
 Ensis macha 11        : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 12        : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 13        : AGTATTGTGTTAACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTGTCCCTCGT
 Ensis macha 14        : AGTATTGTGTTTACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTATCCCTCGT
 Ensis macha 15        : AAGTATTGTGTTTACTTCACTCACCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTAACCCTCGT
 Ensis macha 16        : AGTATTGTGTTTACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTATCCCTCGT
 Ensis macha 17        : AGTATTGTGTTTACTTCACTCACCCGACACTGTATGTTCTTTCTGAGTAGCCAACGTTCGCTCCTTTTAGTTTAAATAGGCATGTGTTAGTATCCCTCGT
 Ensis macha 18        : AAGTATTGTGTTTACTTCACACACCGACACTGTATGTTCTTTCTGAGTAGCCAGCGTTCGCTCCTTTTAGTTTAAATAGGCATGTTTTAGTATCCCTCGT
 Ensis macha 19        : CCCAGGAAAGGCATCTGACGGTTCTGGGCGCAAGGCAGGGCTCTTATGAGCCTACTTCCGGTCCATCTAGTTTCAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 20        : CCCAGGAAAGGCATCTGACGGTTCTGGGCGCAAGGCAGGGCTCTTATGAGCCTACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 21        : CCCAGGAAAGGCATCTGACGGTTCTGGGCGCAAGGCAGGGCTCTTATGAGCCTACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 22        : CCCAGGAAAGGCATCTGACGGTTCTGGGCGCAAGGCAGGGCTCTTATGAGCCTACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 23        : CCCAGGAAAGGCATCTGACGGTTCTGGGCGCAAGGCAGGGCTCTTATGAGCCTACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 24        : CCCAGGAAAGGCATCTGACGGATCTGGGCGCAAGGCAGGGCTCTTATGAGCCAACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 25        : CCCAGGAAAGGCATCTGACGGATCTGGGCGCAAGGCAGGGCTCTTATGAGCCAACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 26        : CCCAGGAAAGGCATCTGACGGATCTGGGCGCAAGGCAGGGCTCTTATGAGCCAACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 27        : CCCAGGAAAGGCATCTGACGGATCTGGGCGCAAGGCAGGGCTCTTATGAGCCAACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 28        : CCCAGGAAAGGCATCTGACGGATCTGGGCGCAAGGCAGGGCTCTTATGAGCCAACTTCCGGTCCATCTAGTTTAAGTAGGCACCTGTTCATGACACTGAT
 Ensis macha 29        : CCCAGGAAAGGCATCTGACGGATCTGGGCGCAAGGCAGGGCTCTTATGAGCCAACTTCCGGTCCATCTAGTTTGAGTAGGCACCTGTTCATGACACTGAT
```

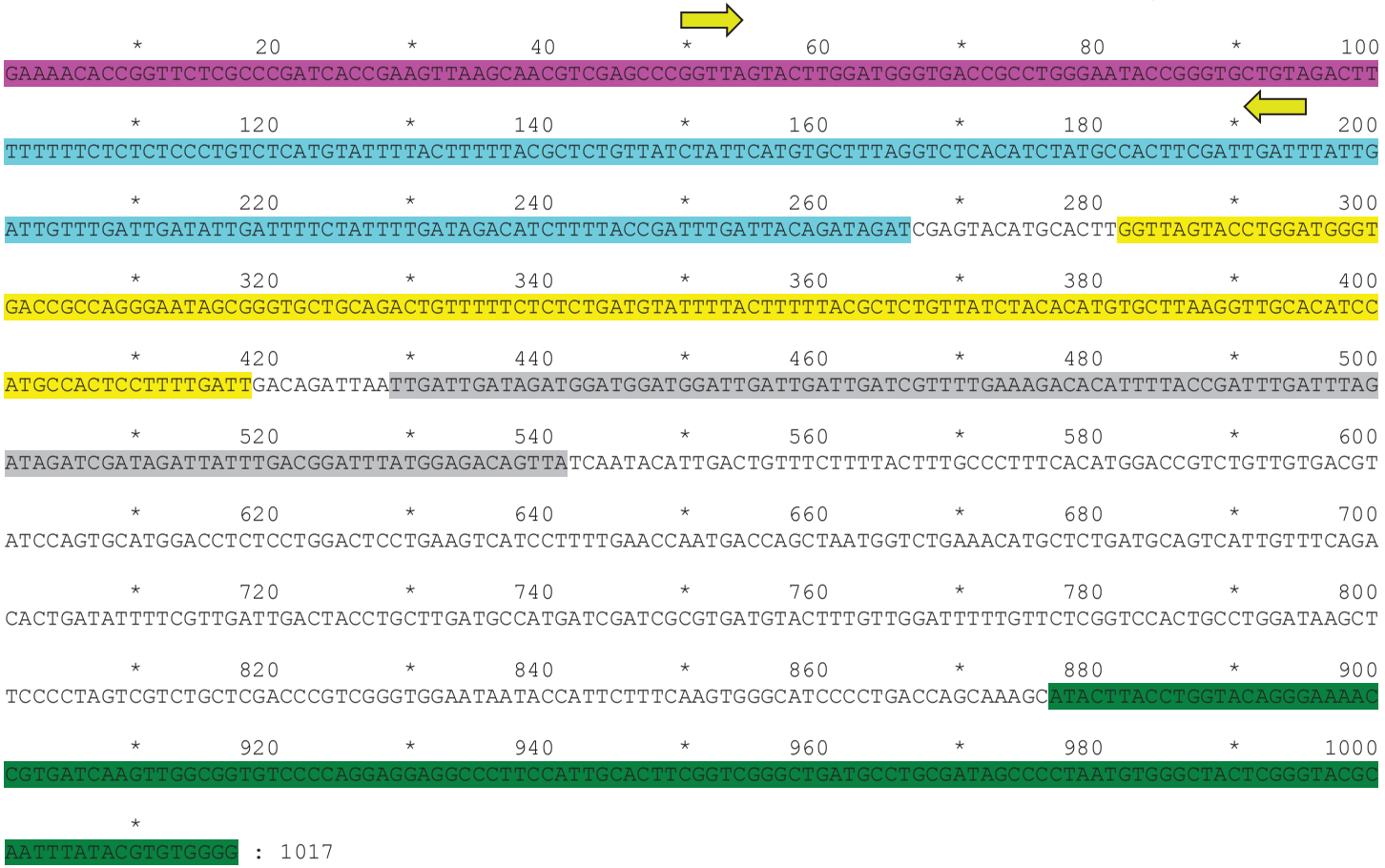

```
 *Ensis macha 30          : AAAGTCAAAGTACGAAATGTGCAGTAAACAACAGCATATAGCTTCGGTATCTTACTTCCGGTCTGTTCAGTTCAAATAGGCCTCTTAAACAGCTTCTGAT
 *Ensis macha 31          : AAAGTCAAAGTACGAAATGTGCAGTAAACAACAGCATATAGCTTCGGTATCTTACTTCCGGTCTGTTCAGTTCAAATAGGCCTCTTAAACAGCTTCTGAT
  Ensis magnus 1          : TTCTTTGCCTGCAAAAGTGTTGTTTTACGTCACGCAGCCGTTCTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAATACGCACATTTAAAGATCGCTCAT
  Ensis magnus 2          : TTCTTTGCCTGCAAAAGTGTTGTTTTACGTCACGCAGCCGTTCTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAATACGCACGTTTAAAGATCGCTCAT
  Ensis magnus 3          : TTCTTTGCCTGCAAAAGTGTTGTTTTACGTCACGCAGCCGTTCTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAATACGCACGTTTAAAGATCGCTCAT
 *Ensis magnus 4          : TAATTAAGGCACAGAAAATGCAGCAAACCACGGCATTTAGCTTCTGTATCTTACTTCCGGTTGTCTTCGGTTCAAATAGTCCTCTCAAACACGTCCTGAT
 *Ensis magnus 5          : TAATTAAAGCACAGAAAGTGCAGCAAACAACGGCATTTAGCTTCTGTAAGTTACTTCCGGTCGTCTTCGGTTCAAATAGTCCTCTGAAACACGTTCTGAT
 *Ensis magnus 6          : TAATTAAAGCACAGAAAGTGCAGCAAACAACGGCATTTAGCTTCTGTAAGTTACTTCCGGTCGTCTTCGGTTCAAATAGTCCTCTGAAACGCGTTCTGAT
  Ensis siliqua 1         : TTCTTTGCCTGCAAAAGTGTTGGTTTACGTCATGCAGCCGTTTTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAATACGCATGTGTAAAGATCGCTCAT
  Ensis siliqua 2         : TCTTTGCCTGCAAAAGTGTTGGTTTACGTCATGCAAGCCGTTTTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAATACGCATGTGTAAAGATCGCTCAT
  Ensis siliqua 3         : TTCTTTGCCTGCAAAAGTGTTGGTTTACGTCATGCAGCCGTTTTGATCAGGCAACGTTCGCTCCTTTCAGTTTAAATACGCATGTGTAAAGATCGCTCAT
  Ensis siliqua 4         : TTCTTTGCCTGCAAAAGTGTTGGTTTACGTCATGCAGCCGCTTTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAATACGCATGTGTAAAGATCGCTCAT
  Ensis siliqua 5         : TCTTTGCCTGCAAAAGTGTTGGTTTACGTCATGCAGCCGTTTTGATCAGCCAAACGTTCGCTCCTTTCAGTTTACATACGCATGTGTAAAGATCGCTCAT
  Ensis siliqua 6         : TTCTTTGCCTGCTAAAGTGTTGGTTTACGTCATGCAGCCGTTTTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAATACGCATGTGTAAAGATCGCTCAT
*#Ensis siliqua 7         : TAATTAAAGCACACAAAGTGCAGCAAACAACGGCATTTAGCTTCTGTATCTTACTTCCGGTCGTCTTCGGTTCAAATAGTCCTCTCAAACACGTCCTGAT
 *Ensis siliqua 8         : TAATTAAAGCACACAAAGTGCAGCAAACAACGGCATTTAGCTTCTGTATCTTACTTCCGGTCGTCTTCGGTTCAAATAGTCCTCTCAAACACGTCCTGAT
 #Ensis ensis 1           : TTCTTTGCCTGCAAAAGTGTTGTTTTACGTCACGCAGCCGTTCTGCTCAGCCAACGTTCGCTCCTTTCAGGTTAAAAACGCACATTTAAAGATCGCTGTT
 *Ensis ensis 2           : TAATTAAAGCACAGAAAGTGCAGCAAACAACGGCATTTAGCTTCTGTACCTTACTTCCGGTCGTCTTCGGTTCAAATAGTCCTCTGAAACACGTTCTGAT
 *Ensis ensis 3           : CAATTAAAGCACAGACAGTGCAGCAAACAACGGCATTTAGCTTCTGTACCTTACTTCCGTTCGTCTTCGGTTCAAATAGTCCTCTGAAACACGTTCTGAT
  Ensis ensis 4           : TGCGTCTTTGCGTACAAAGTGTGTTTACGTCACGCAGCCGTTCTGATCAGCCAACGTTCGCTCCTTTCAGTTTAAAAACGCACGTTTAAAGATCGCTCAT
  Ensis goreensis 1       : CCTGCAAAAGTGTTGTTTTACGCCACTCACCCCGGCGTTCTGATCAGCCAACGACGTTCGCTCCTTTCAGTTTAAGTACGCACGTGTAAAGATCTCTTAT
 *Ensis minor Chenu 1     : TAATTAAAGCACACAAAGTGCAGCAAACAACGGCATTTAGCTTCTGTATCTTACTTCCGGTCGTCTTCGGTTCAAATAGTCCTCTCAAACACGTCCTGAT
 *Ensis minor Chenu 2     : TAATTAAAGCACACAAAGTGCAGCAAACAACGGCATTTAGCTTCTGTATCTTACTTCCGGTCGTCTTCGGTTCAAATAGTCCTCTCAAACACGTCCTGAT
 *Ensis minor Dall 1      : AAAATTGAAGTACGAAATGTGCAGTAAACAACAGCATATAGCTTCAGTATCTTACTTCCGGTCTGTTCAGTACAAATAGTCCTCTTAAACCACTTCTGAT
  Pharus legumen 1        : TGACCAGTGGGTTCACAGCAGTCTCCACGGGCTGGTGCAACTTACTGTCAACACTTCCGGTTTATGGACGTTTAAATAGCGATGTTACAAAGCACCTAAT
 *Pharus legumen 2        : ACTGAAATACTGTTGGTGTAGACGTAAAGCAACAGTCAAACAAGCAAGCAAACACACAAACAGTTTCATACATAAGAAGGCAGTCGATTCAGCCCTTAAT
 *Pharus legumen 3        : ACTGAAATACTGTTGGTGTAGACGTAAAGCAACAGTCAAACAAGCAAGCAAACACACAAACAGTTTCATACATAAGAAGGCAATCGATTCAGCCCTTAAT
 *Pharus legumen 4        : TCGAAATTGAAAGCTGGCACGATCATAATCCTGATCCAAATGCTGAGAACACGGGGCCAGGTGTGTACCCTTTAAATAGCGCTCGTGGCGATCATTCAAC
 *Pharus legumen 5        : TCGAAATTGAAAGCTGGCACGATCATGATCCTGATCCAAATGCTGAGAACACGGGGCCAGGTGTGTACCCTTTAAATAGCGATCGTGGCGATCATTCAAC
 *Pharus legumen 6        : TCGAAATTGAAAGCTGGCACGATCATGATCCTGATCCAAATGCTGAGAACACGGGGCCAGGTGTGTACCCTTTAAATAGCGCTCGTGGCGATCATTCAAC
 *Ensiculus cultellus 1 : AATTATTTATCAAGTCATTTACTTCTTTACCTATTTGAAGAATTTGCCAGGCAACTTCCGGTCTGATCGATTGAAATAGCCCTTTCAAAGCCACCCTCAT
 *Ensiculus cultellus 2 : AGGATGGCAGGTAAATGATGCTTCGTACTATCCAGGTCGGTTGGGACCCTACCGCTTCCGTTATACCCAGTTCAAATGGACATGTGCAGCAGCTACCTTT
 *Ensiculus cultellus 3 : AGGATGGCAGGTAAATGATGCTTCGTACTATCCAGGTCGGTTGGGACCCTACCACTTCCGTTATACCCAGTTCAAATGGACATGTGCAGCAGCTACCTTT
 *Siliqua patula 1       : TACAAGCATACATACATTCATTTATTAAGGTGTTTGTTGTACATCAGTAGCCTGCATCCGCTTCCAACAGTATAAATACACCTTATAACGGAGCACTTTC
 *Siliqua patula 2       : CAATAATATGCACAGAGCCCCAATACGGATCACCGAAGTCATCATTGTGTCAGGCAGCCGGTTCCAGCCACTTAAATTGCCCGTTTAACAGGGAACTTTC
 *Siliqua patula 3       : ACGCATTAATATGCACAGAGCCCAATTACAGATAGAAGTCAGCATTGTGTCTGGCAGCCGGTTCCAACCGTTTAAATTGCCCTTTTAACAGGGAACTTTC

  Ensis magnus 1          : GCCATTTCAGAGGTGTGGAGTCATTCGAAAAAAACAATGCACAACACGCCGGGTCGAATAACGACAGCTTCTCAAAAGGGTTGCGCATGCATGGCAAAGC
  Ensis magnus 2          : GCCATTTCAGAGGTGTGGAGTCATTCGAAAAAATCAATGCACAACACGCCGGGTCGAATAACGACAGCTTCTCAAAAGGGTTGCGCATGCCTGGCAAAGC
  Ensis magnus 3          : GCCATTTCAGAGGTGTGGAGTCATTCGAAAAAAACAATGCACAACACGCCGGGTCGAATAACGACAGCTTCTCAAAAGGGTTGCGCATGCATGGCAAAGC
  Ensis ensis 1           : GCCATTTCAGAGGTGTGGAGTCATTCGAAAAAATCAATGCACAACACGCCGGGTCGAATAACGGCAGCTTCTCAAAAGGGTTGCGCATGCCTGGCAGAGC
  Ensis ensis 2           : GCCATTTCAGAGGTGTGGAGTCATTCGAAAAAACCAATGCACAACACGCCGGGTCGAATAACGACAGCTTCTCAAAAGGGTTGCGCATGCCTGGCAAAGC
  Ensis minor Chenu 1     : GCCATTTCAGATGTCTGGAGTCATTCGAAAAAACAAATGCACAACACGCCGGGTCGAATAACGACAGCTTCTCAAAAGGGTTGCGCCTGCATGCCAAAGC
  Ensis minor Chenu 2     : GCCATTTCAGATGTCTGGAGTCATTCGAAAAAACAAATGCACAACACGCCGGGTTGAATAACGACAGCTTCTCAAAAGGGTTGCGCCTGCATGGCAAAGC
  Ensis siliqua 1         : GCCATTTCAGATGTCTGGAGTCATTCGAAAAAACAAATGCACAACACGCCGGGTCGAATAACGACAGCTTCTCAAAAGGGTTGCGCCTGCATGGCAAAGC
  Ensis siliqua 2         : GCCATTTCAGATGTCTGGAGTCATTCGAAAAAACAAATGCACAACACGCCGGGTCGAATAACGACAGCTTCTCAAAAGGGTTGCGCCTGCATGGCAAAGC
  Ensis directus          : TAAATAGCAGTGGTGTGGGGTCAATTCCAAAGGTCACTGTCCGACCTGTCGGGCCGAAAAACAACGTTGTGTGAAAAGGTATGCACGTGCATTGGAAGGC
  Ensis macha 1           : TGAAAAATCAGTGGTGTGGAGTCAATTCGAAAATCACTGTCCGACACGTCGGGCCGAATAATACCCTTGTGTGAAAAGGTATGCACGTGCATAGGAAGGC
```

```
→Ensis macha 2          : CTCGGTCCACTGCCTGGATAAGCTTCCCCTAGTCGTCTGCTCGACCCGTCGGGTGGAATAATACCATTCTTTCAAGTGGGCATCCCCTGACCAGCAAAGC
 Ensis macha 3          : TGAAAAATCAGTGGTGTGGAGTCAATTCGAAAATCACTGTCCGACACGTCGGGCCGAATAATACCCTTGTGTGAAAAGGTATGCACGTGCATAGGAAGGC
#Ensis minor Dall       : TAAATAGCAGTGGTGTGGAGTCAATCGCAGAGATCACTATACGACGTGTCGGGCCGAATAACAACGTTGTTTGAAAAGGTATGCACGTACATAGGAAGGC
 Pharus legumen 1       : AATGACACAAACAGCCGATTGGACGTGTTCCAACGTCTATCAACCTGCCTGTGGGAAAGACTATCCTCCATACAAATGAGCTGTGTGAGCGGAACAACTC
 Pharus legumen 2       : AATGACACAAACAGCCGATTGGACGTGTTCCAACGTCTATCAACCTGCCTGTGGGAAAGACTATCCTCCATACAAATGAGCTGTGTGAGCGGAACAACTC
 Pharus legumen 3       : AATGACACAAACAGCCGATTGGACGTGTTCCAACGTCTATCAACCTGCCTGTGGGAAAGACTATCCTCCATACAAATGAGCTGTGTGAGCGGAACAACTC
#Siliqua patula 1       : AAGTCGACCTCAACGCCACCCTGGATGCGCTGCCCCATGTAAACCTGTTCTTGCGTTTACCTGGCGCCAATACGTAAGGTTCCCATCATGTAGGCAAAGC
#Siliqua patula 2       : AAGTCGACCTCAACGCCACCCTGGATGCGCTGCCCCATGTAAACCTGTTCTTGCGTTTACCTGGCGCCAATACGTAAGGTTCCCATCATGTAGGCAAAGC
 Ensiculus cultellus 1  : TTCGTGCATGTGTAGCAACAAAACCCCTCAGTGCGCTAAAATCACCTAGCAGACCGAAAAATGTTCTCTTGATAAAAGATCAACTAATATGCCCCCAATC
 Ensiculus cultellus 2  : TTCGTGCATGTGTAGCAACAAAACCCCTCAGTGCGCTAAAATCACCTAGCAGACCGAAAAATGTTCTCTTGATAAAAGATCAACTAATATGCCCCCAATC
 Ensiculus cultellus 3  : TTCGTGCATGTGTAGCAACAAAACCCCTCAGTGCGCTAAAATCACCTAGCAGACCGAAAAATGTTCTCTTGATAAAAGATCAACTAATATGCCCCCAATC
 Aplysia californica    : TCCATTGCACTCCGGTATGGCTGACCCCTGCGATCACTAAATTGGTGACTCGAGTGCGTAATTTTTCCGTAGGGGGGACTGCGTTCGCGCTATCCCCTGA
 Lottia gigantea        : TTGCACCTAGCGGAGGCTGACCCCTGCGATCACCCCTAATGTGGGTGACTCGAGTACGTAATTTTTTAGTATGGGGGACTGCGTTCGCGCTCTCCCCTGG
```

Supplementary File S3 Conserved DNA motifs found within the upstream regions. Positions relative to the transcription start sites are specified. (a) Motif upstream the 5S ribosomal DNA transcription start site. We included sequences from all the razor shell species studied in this survey (see main text). (b) Motif upstream the U1 small nuclear DNA transcription start site. We included sequences from all the razor shell species studied in this survey (see main text), and from the gastropods *Aplysia californica* and *Lottia gigantea*. Logos constructed using WebLogo 2.8.2 available at http://weblogo.berkeley.edu/.

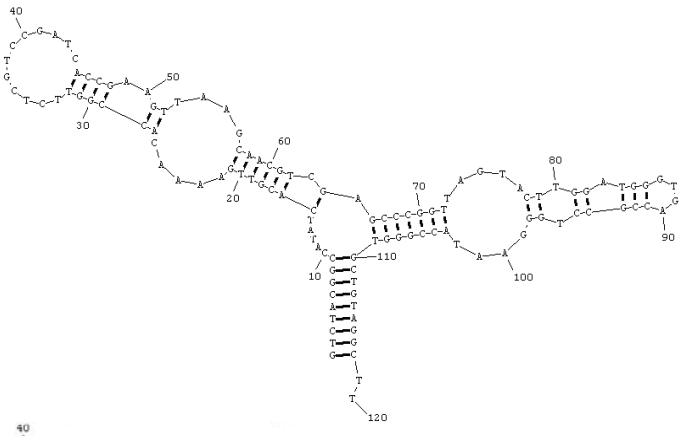


Supplementary File S4 DNA sequence of an *Ensis macha* 5S-U1 clone. The pink region corresponds to the RNA coding region of 5S ribosomal DNA (5S). Region shaded in green is a partial RNA coding region of U1 small nuclear DNA (type A, see main text). The nontranscribed spacer (NTS) located between both RNA coding regions is referred to as iota NTS in the main text. Light blue shaded region is similar to epsilon 2 NTS from the species *E. macha* (see main text). Yellow region is a truncated pseudogenised 5S rDNA copy, as it is similar to the 5S and NTS flanked by the arrows. The grey region is similar to a DNA fragment associated to the *Taenia solium* spliced leader sequence and spliced leader mini-exon (accession number AJ428456; identities=72%, gaps=11%, E value=$4 \cdot 10^{-6}$).

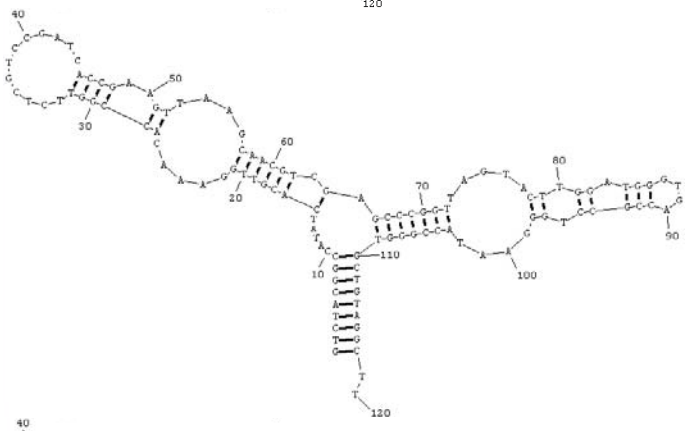


Supplementary File S5 5S ribosomal RNA predicted secondary structures of razor shells. Sequences used were complete after excluding the primer-annealing regions. Structures correspond to one or more species and species can have one or more structures. The downstream nontranscribed spacer group(s) of each 5S ribosomal coding region is indicated in parentheses (see main text). (a) *Ensis directus* (alpha). (b) *E. directus* (alpha). (c) *E. directus* (gamma, delta). (d) *E. macha* (theta). (e) *E. magnus*, *E. minor* (Chenu) (eta). (f) *E. ensis* (eta). (g) *Siliqua patula* (pi, rho); *E. directus* (alpha). (h) *S. patula* (omicron). (i) *Ensiculus cultellus* (nu). (j) *Pharus legumen* (lambda). (k) *P. legumen* (lambda). (l) *E. ensis* (zeta). (m) *E. directus* (alpha). (n) *E. siliqua* (eta). l, m, and n are putative pseudogenised copies (see main text).

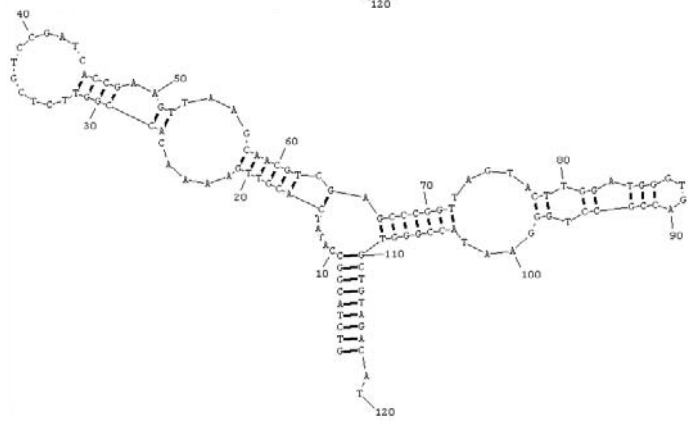
a



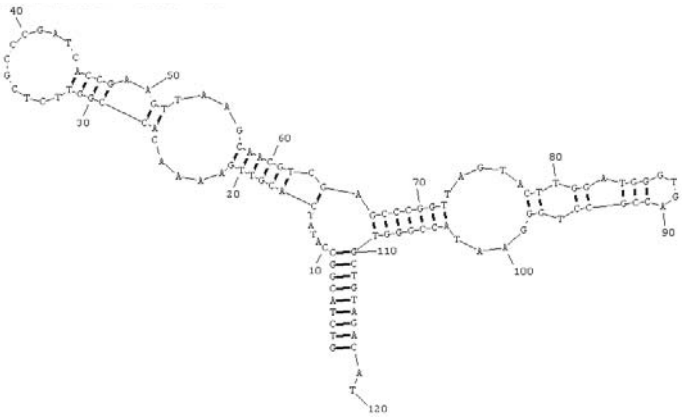
b



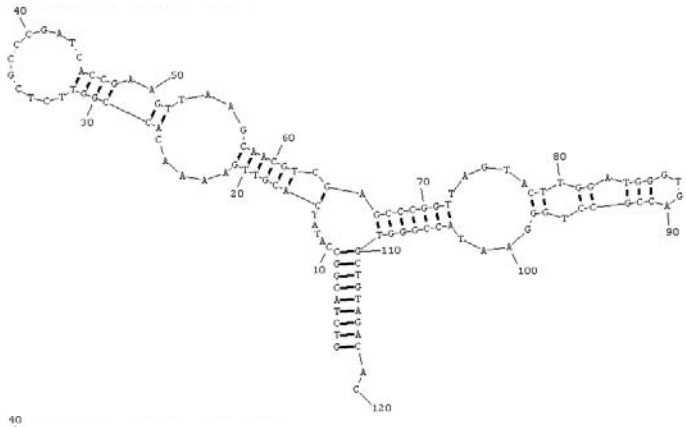
c



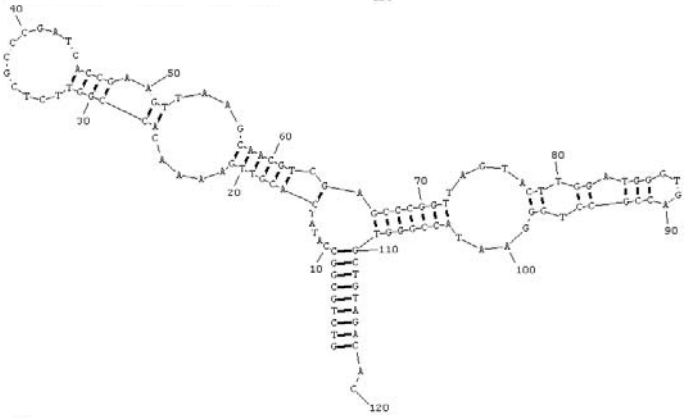
d



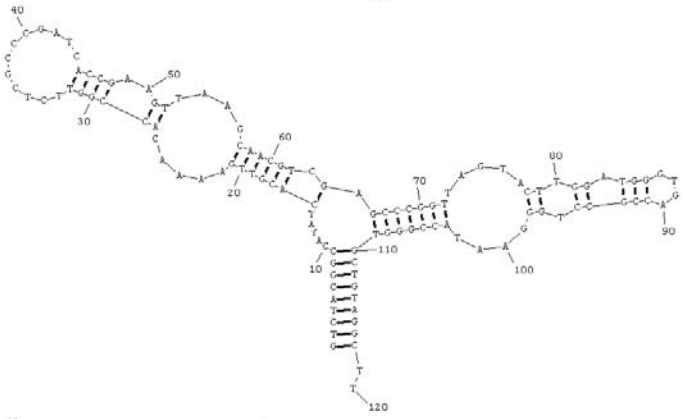
e



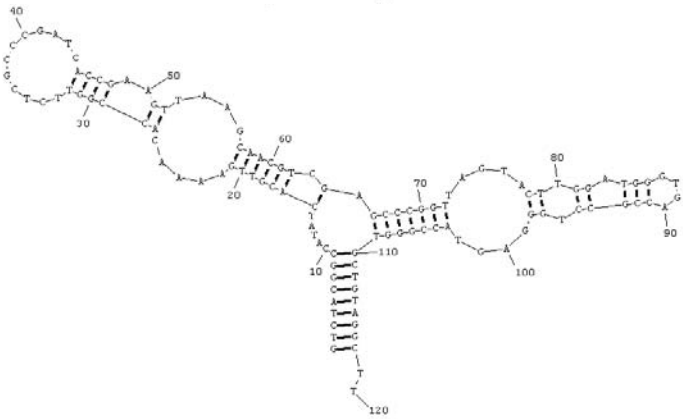
f



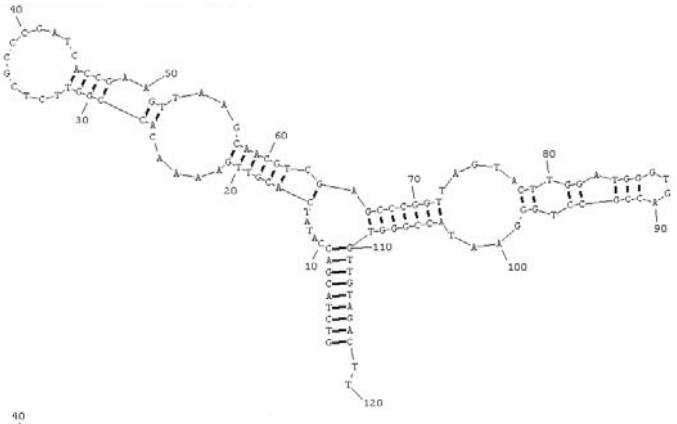
g



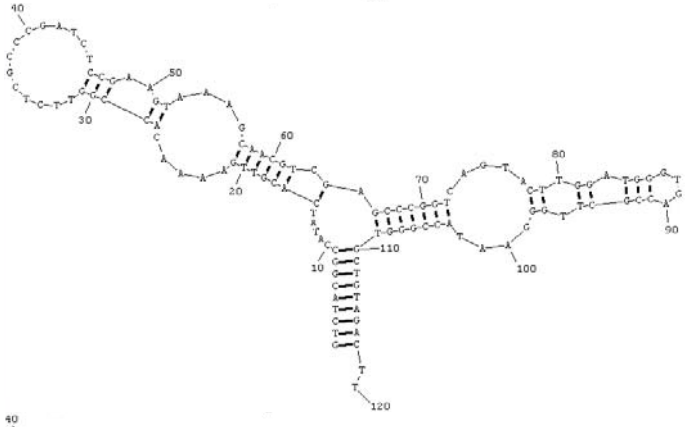
h



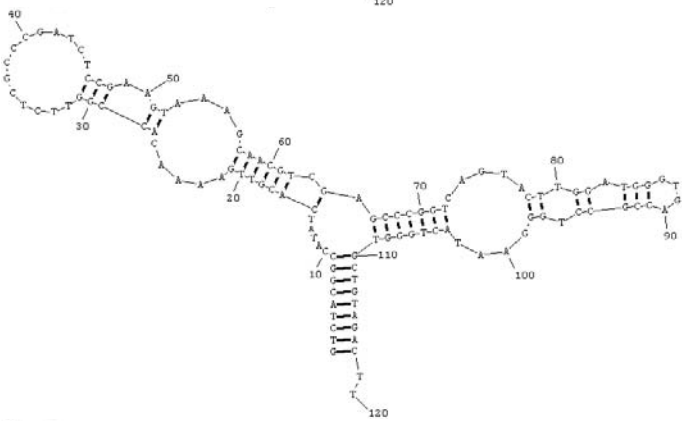
i



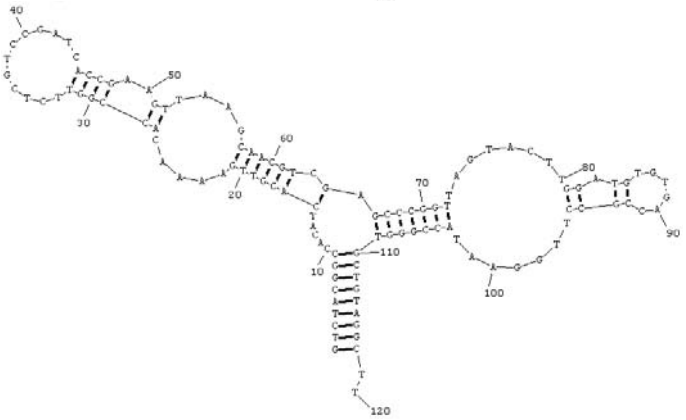
j



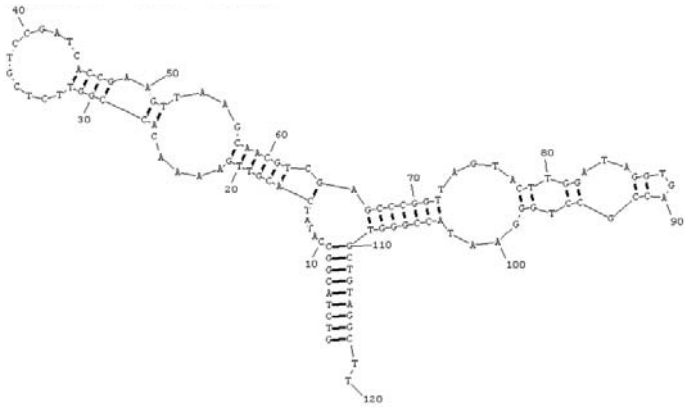
k



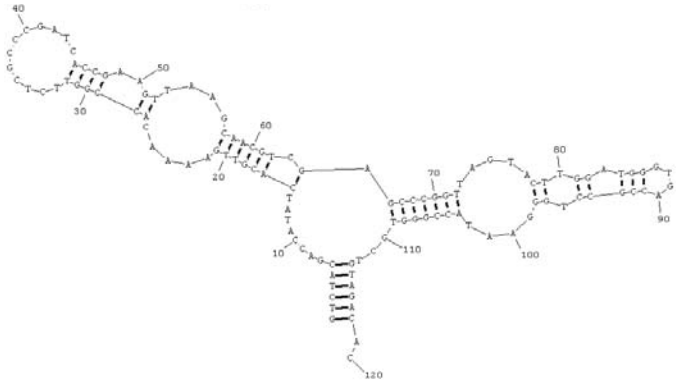
l



m

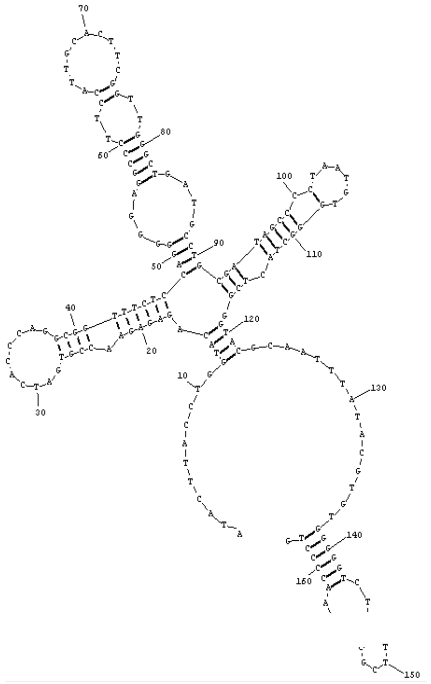

40
50
30
60
20
70
80
90
10
110
100
120

n

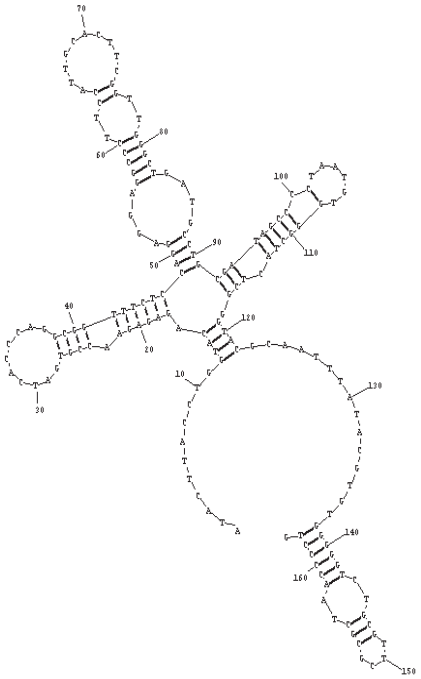

40
50
30
60
20
70
80
90
10
110
100
120

Supplementary File S6 U1 small nuclear RNA predicted secondary structures of razor shells and gastropods. (a) *Ensis minor* (Chenu)[1]. (b) *E. magnus*[2], *E. ensis*[1], *E. siliqua*[1]. (c) *E. macha* type A[3]. (d) *E. macha* type B[1]. (e) *E. directus*[3]. (f) *Ensiculus cultellus*[1]. (g) *Pharus legumen*[2]. (h) *Aplysia californica*[2]. (i) *Lottia gigantea*[2]. (j) *E. minor* Dall[3]. (k) *Siliqua patula*[1]. ([1]) sequence built up with two clones from the same individual; ([2]) complete sequence after excluding the primer-annealing regions; ([3]) sequence completed with the last 23 nucleotides from *E. magnus* U1 small nuclear RNA (see main text); j and k are putative pseudogenised copies (see main text).

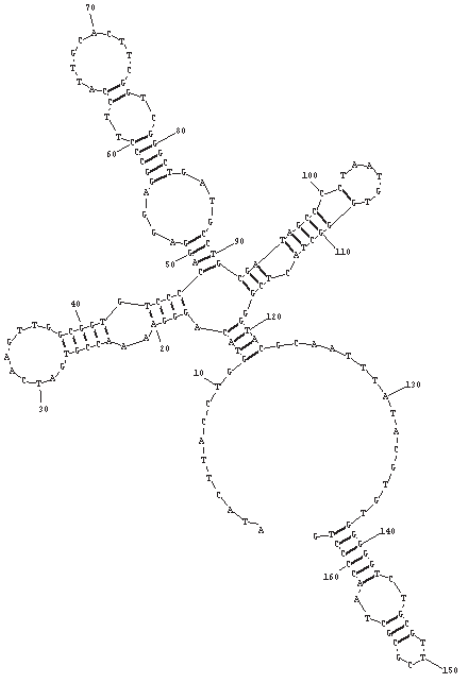
a



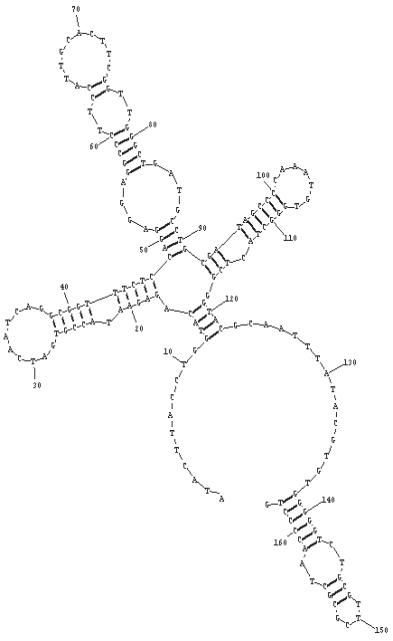
b



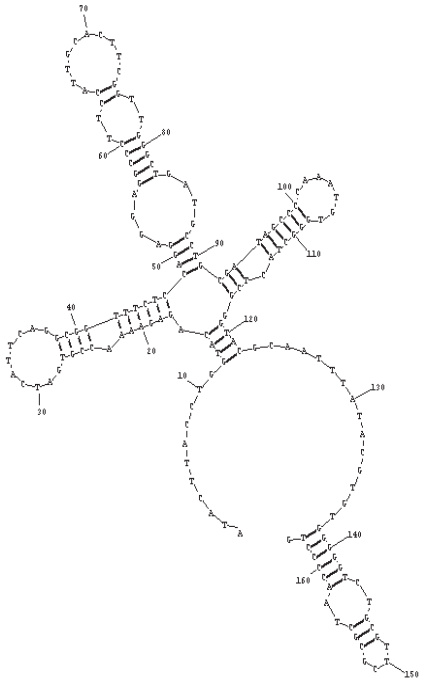
c



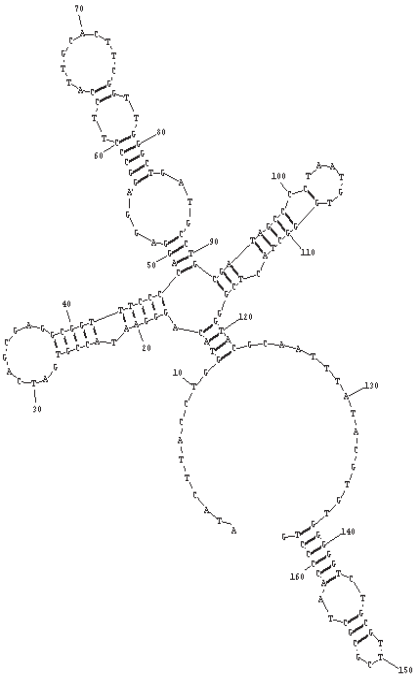
d



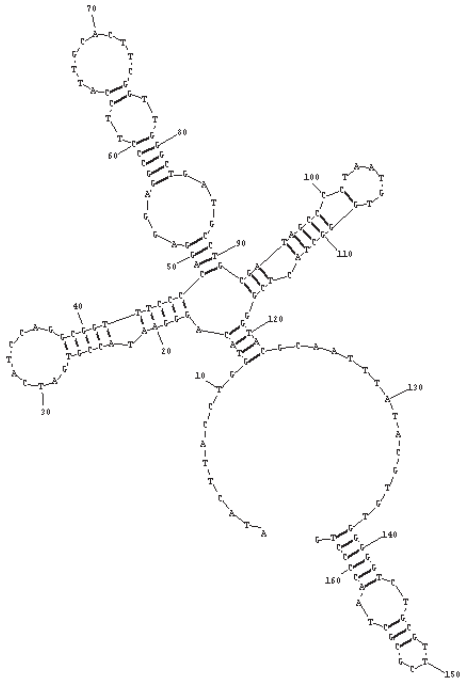
e



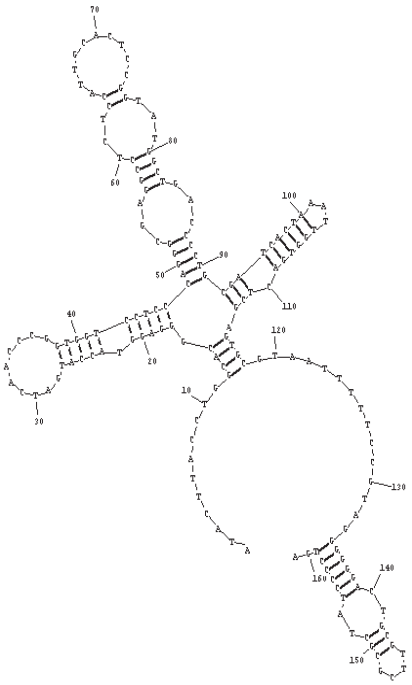
f



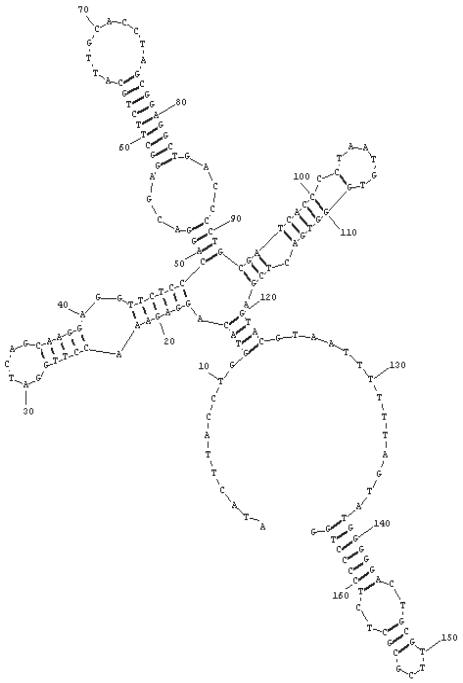
g



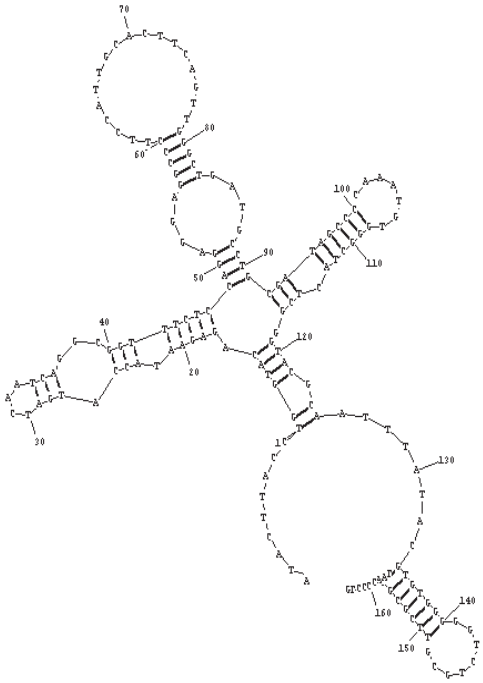
h



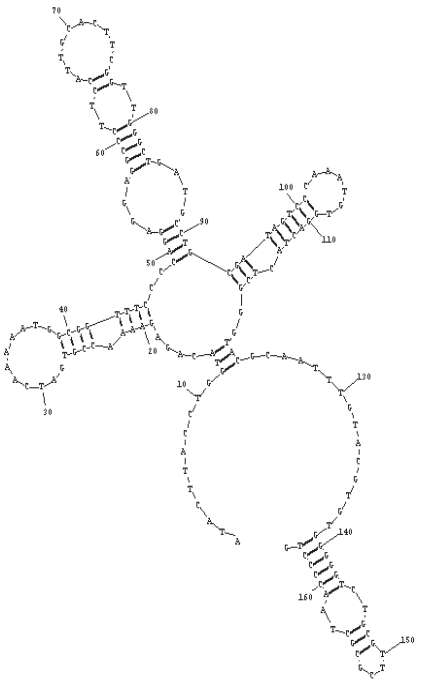
i



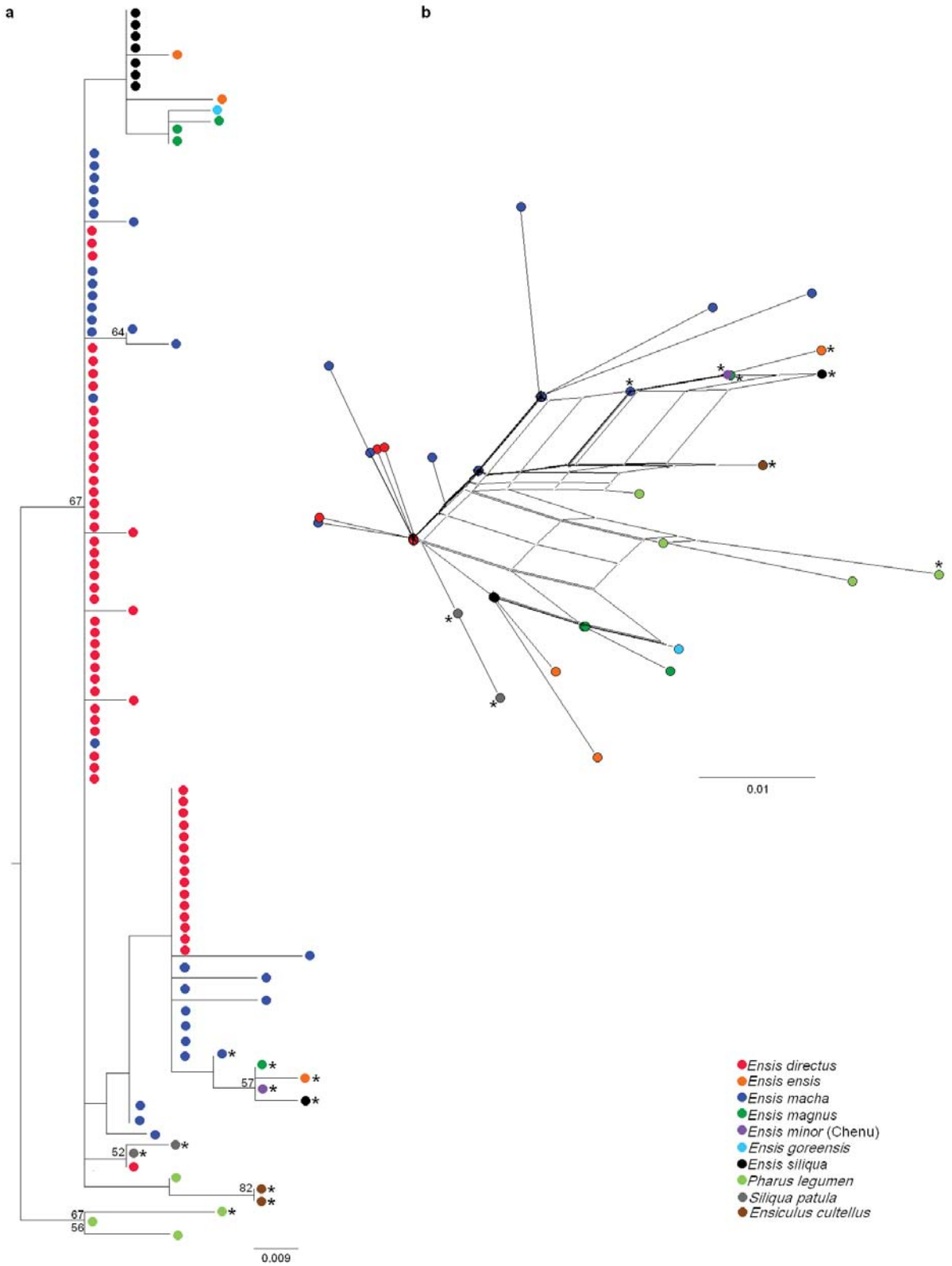
j



k



Supplementary File S6 Phylogenetic relationships of 5S ribosomal RNA coding sequences from razor shells species. Sequences are not clearly clustered by species, and the low bootstrap values obtained do not support a clustering according to gene variants either. Asterisks (*) indicate the copies retrieved from mixed clones of 5S ribosomal DNA and U1 small nuclear DNA. (a) Maximum likelihood phylogenetic tree constructed using the K80 model. Numbers on the tree correspond to nonparametric bootstrap supports (1000 replicates) and they are reported only for nodes with values ≥50. (b) Phylogenetic network constructed using the neighbor net algorithm. For NTS types, see Table 3.



4.1.2 Systematic analysis and evolution of 5S ribosomal DNA in metazoans

Joaquín Vierna, Stefanie Wehner, Christian Höner zu Siederdissen, Andrés Martínez-Lage, Manja Marz (2013) Systematic analysis and evolution of 5S ribosomal DNA in metazoans. Heredity 111:410-421.

Bibliometrics 2012 JCR Science Edition

Impact factor: 4.110

Ecology: Q1

Evolutionary Biology: Q2

Genetics & Heredity: Q1

Heredity (2013) 111, 410–421

www.nature.com/hdy

ORIGINAL ARTICLE

Systematic analysis and evolution of 5S ribosomal DNA in metazoans

J Vierna[1,2,6], S Wehner[3,4,6], C Höner zu Siederdissen[5], A Martínez-Lage[1] and M Marz[3,4]

Several studies on 5S ribosomal DNA (5S rDNA) have been focused on a subset of the following features in mostly one organism: number of copies, pseudogenes, secondary structure, promoter and terminator characteristics, genomic arrangements, types of non-transcribed spacers and evolution. In this work, we systematically analyzed 5S rDNA sequence diversity in available metazoan genomes, and showed organism-specific and evolutionary-conserved features. Putatively functional sequences (12 766) from 97 organisms allowed us to identify general features of this multigene family in animals. Interestingly, we show that each mammal species has a highly conserved (housekeeping) 5S rRNA type and many variable ones. The genomic organization of 5S rDNA is still under debate. Here, we report the occurrence of several paralog 5S rRNA sequences in 58 of the examined species, and a flexible genome organization of 5S rDNA in animals. We found heterogeneous 5S rDNA clusters in several species, supporting the hypothesis of an exchange of 5S rDNA from one locus to another. A rather high degree of variation of upstream, internal and downstream putative regulatory regions appears to characterize metazoan 5S rDNA. We systematically studied the internal promoters and described three different types of termination signals, as well as variable distances between the coding region and the typical termination signal. Finally, we present a statistical method for detection of linkage among noncoding RNA (ncRNA) gene families. This method showed no evolutionary-conserved linkage among 5S rDNAs and any other ncRNA genes within Metazoa, even though we found 5S rDNA to be linked to various ncRNAs in several clades.
Heredity (2013) **111,** 410–421; doi:10.1038/hdy.2013.63; published online 10 July 2013

Keywords: 5S rRNA; homologous genes; noncoding RNA; secondary structure; paralogs; birth-and-death evolution

INTRODUCTION

The evolution of 5S ribosomal DNA (5S rDNA) has been studied in some groups of organisms, mainly within genera or within families (for example, Martins and Wasko (2004); Rooney and Ward (2005); Vierna *et al.* (2009, 2011); Freire *et al.* (2010); Perina *et al.* (2011); Vizoso *et al.* (2011)). Nevertheless, several intriguing features, such as high conservation along evolution in contrast to high intragenomic divergence, a plastic genomic organization and linkage to other genes, make this multigene family an interesting issue in evolutionary genetics that deserves a large-scale analysis.

5S rDNA (as well as other ribosomal genes) is expected to display low intragenomic divergence levels owing to the occurrence of homogenizing mechanisms (unequal crossing-overs and gene conversions) that are favored by the tandem arrangement of these genes and lead to so-called concerted evolution (reviewed in Eickbush and Eickbush (2007)). However, many reports have been recently published in which the concerted evolution model did not explain the intragenomic divergence found in some organisms, mainly (but not exclusively) within the non-transcribed spacer (NTS) region (Rooney and Ward, 2005; Fujiwara *et al.*, 2009; Vierna *et al.*, 2009, 2011; Freire *et al.*, 2010; Úbeda-Manzanaro *et al.*, 2010; Perina *et al.*, 2011; Vizoso *et al.*, 2011). Other evolutionary models (birth-and-death evolution; mixed process of concerted and birth-and-death evolution (Nei and Rooney, 2005)) have been proposed to drive the evolution of 5S rDNA (Rooney and Ward, 2005; Fujiwara *et al.*, 2009; Vierna *et al.*, 2009, 2011; Freire *et al.*, 2010; Úbeda-Manzanaro *et al.*, 2010; Perina *et al.*, 2011; Vizoso *et al.*, 2011).

5S rDNA is present in a variable number of repeats (usually, hundreds of copies) in each genome. These repeats can occur in tandem forming long arrays in some species, whereas in other cases they are dispersed throughout the genome. In some organisms, 5S rDNA repeats have been found linked to other noncoding RNA (ncRNA) gene families, such as small nuclear RNAs (snRNAs) (Vahidi *et al.*, 1988; Nilsen *et al.*, 1989; Zeng *et al.*, 1990; Keller *et al.*, 1992; Pelliccia *et al.*, 2001; Cross and Rebordinos, 2005; Manchado *et al.*, 2006; Marz *et al.*, 2008; Freire *et al.*, 2010; Vierna *et al.*, 2011; Vizoso *et al.*, 2011) or to protein-coding genes such as histones (Eirin-Lopez *et al.*, 2004).

Although linkages of 5S rDNA to other ncRNAs have been shown also in bacteria (Gongadze, 2011), protists (Drouin and Tsang, 2012) and plants (Wicke *et al.*, 2011; Layat *et al.*, 2012) for longer time scales, the animal linkages of ncRNAs seem not to be stable over long

[1]Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, A Coruña, Spain; [2]AllGenetics, Ed. de Servicios Centrales de Investigación, Campus de Elviña s/n, A Coruña, Spain; [3]RNA Bioinformatics Group, Department of Pharmaceutical Chemistry, Philipps-Universität Marburg, Marbacher Weg 6, Marburg, Germany; [4]Department for Bioinformatics, Faculty of Mathematics and Computer Science, Friedrich-Schiller-University of Jena, Jena, Germany and [5]Department of Theoretical Chemistry, University of Vienna, Vienna, Austria
Correspondence: Professor M Marz, Department of Bioinformatics, Faculty of Mathematics and Computer Science, Friedrich-Schiller-University of Jena, Leutragraben 1, Jena D-07742, Germany.
E-mail: manja@uni-jena.de
[6]Joint first authors.
Received 23 September 2012; revised 9 April 2013; accepted 17 May 2013; published online 10 July 2013

evolutionary time scales. They appear to be the result of stochastic processes within genomes with no effect on fitness, even though this has not been demonstrated (see Drouin and Moniz de Sá (1995) for a review). Interestingly, 5S rDNA repeats can show different organization modes in the same species (Little and Braaten, 1989), and their transposition could be frequent within genomes, as reported by Drouin and Moniz de Sá (1995); Kalendar *et al.* (2008); Cohen *et al.* (2010).

Reports on the evolution of 5S rDNA in various animal and fungi groups have been published during the last few years, and all (Martins and Wasko, 2004; Vierna *et al.*, 2009, 2011; Freire *et al.*, 2010; Úbeda-Manzanaro *et al.*, 2010; Perina *et al.*, 2011; Vizoso *et al.*, 2011) except one (Rooney and Ward, 2005) have relied on data obtained from PCR-cloning-sequencing techniques. Even though these procedures are appropriate when working with non-model organisms, they may not give a complete picture of the features and diversity of this multigene family. Fortunately, this can be solved by using genome project data, when available. Here, we have obtained a huge set of animal 5S rDNA candidate sequences, which were carefully filtered according to stringent criteria. Additionally, we gathered a set of U1 small nuclear DNA sequences from the same metazoan genomes, that were used in the linkage analysis between 5S rRNA and other ncRNAs.

MATERIALS AND METHODS

Sequence data

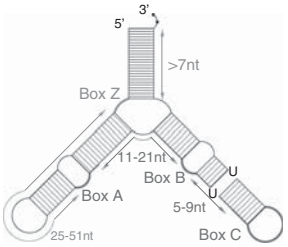
Previously known 5S rRNA and U1 snRNA sequences were taken from Rfam (Gardner *et al.*, 2011) and selected previous studies (Marz *et al.*, 2008; Vierna *et al.*, 2009, 2011). These sequences (available from the electronic supplement http://www.rna.uni-jena.de/supplements/5SRNA/index.html) were used as an initial query in the development of a candidate pool (see below). The source, composition, download dates, assembly status, coverage, real number of nucleotides and expected number of nucleotides (from the animal genome size database (Gregory, 2012)) of all genomes analyzed are listed in the electronic supplement as well.

Homology search for 5S rRNAs and U1 snRNAs

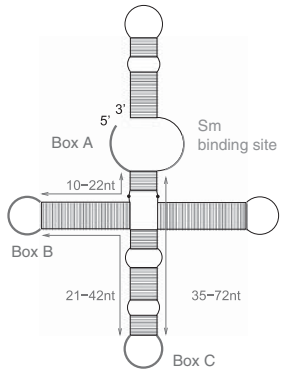
Development of a candidate pool. First, we used blast (Altschul *et al.*, 1990) with a low *E*-value $<10^{-4}$ to get as many 5S rRNA and U1 snRNA candidates as possible. Overlapping hits were merged and extended 50 nt in both directions, manually viewed using emacs ralee mode (RNA ALignment Editor in Emacs) (Griffiths-Jones, 2005) and cut into their expected length. Consensus sequences of each alignment block and species were added to the query data set. We repeated this blast search with the same parameters and the collection step for all organisms until no new reliable candidates were found.

Sequence conservation. After having studied in detail previously reported 5S rRNA and U1 snRNA sequences, we selected four conserved motifs in animals for each ncRNA, Figure 1. Subsequently, we wrote rnabob descriptors, which characterized the conserved motifs (boxes and Sm-binding site) and their allowed distances (Figure 1, right). We decided against a covariance model, as we did not want a high variability and speed up the analysis. To detect divergent 5S rRNAs, we allowed point mutations to occur in one of the boxes, and variable distances between motifs. Additionally, box Z for the 5S rRNA and Sm-binding site for U1 snRNA were used in six species because of its huge initial candidate set: *Homo sapiens*, *Pongo pygmaeus*, *Macaca mulatta*, *Bos taurus*, *Pteropus vampyrus* and *Saccoglossus kowalevskii*. Candidates that did not fulfill these criteria were not discarded but marked with a demerit for further analysis.

Structure conservation. In a second step, we examined the secondary structure of the candidates. If RNAfold (Hofacker, 2003) did not fold the sequences



Organisms	Box A	Box B	Box C	Box Z	Lit.
Consensus	YAAGCA	UUAGUA	GGGWGA		
Vertebrates	U.....	.....	...A..		Martins and Wasko (2004)
Xenopus sp.	C.....	.....	...A..		Pieler *et al.* (1987)
Bombyx mori	C.....	.C....	...U..		Komiya *et al.* (1981)
Honey bee	C....U	.....	...U..		Gillespie *et al.* (2006)
Drosophila sp.	U.....	.....	...G..		Sharp and Garcia (1988)
Razor Shells	U.....	.N....	...U..		Vierna *et al.* (2011)
Plasmodium	U.....	C.....	...A..		Shippen-Lentz and Vezza (1988)
Search motif	-AAGC-	--AGUA	GGGNGA	NNNNNCAUA	
Mutations	1	1	1	1	



Organisms	Box A	Box B	Box C	Sm bind. site	Lit.
Consensus	AWACUUACCUG	GAUCA	AUUGCAC	UAAUUUNUG	
Vertebrates	.U.....	.....	.....	.....G..	Marz *et al.* (2008)
"Basal Deut."	.U.....	.....	.....	.....C..	Marz *et al.* (2008)
Insects	----.....	.....	.....	.....U..	Marz *et al.* (2008)
A. aquaticus	.U.....	.....	.....	.....C..	Barzotti *et al.* (2003)
"worms"	-A.....	.....	.....	.....U..	Marz *et al.* (2008)
C. elegans	.A.....	.....	.....	.....U..	Thomas *et al.* (1990)
A. lumbricoides	.A.....	.....	.....	.....U..	Shambaugh *et al.* (1994)
Trypanosoma	.U.....	.....	.....	.....G..	Schnare and Gray (1999)
Razor Shells	.U.....	.....	.....	C.....A..	Vierna *et al.* (2011)
Tetrahymena	--.....	.....	.....	.....C..	Ørum *et al.* (1991)
Search motif	----UUACCUG	GAUCA	AUUGCAC	-AAUUUNU-	
Mutations	2	2	2	1	

Figure 1 Known conserved motifs of 5S rRNA (top) and U1 snRNA (bottom) according to previous reports, and motifs used in the filtering step. For each motif, a specific number of mutations was accepted. Allowed distances between motifs are displayed on the left. Red motifs were used to filter sequence candidates in all species. Green motifs were used as additional filters in some species (*H. sapiens*, *P. pygmaeus*, *M. mulatta*, *B. taurus*, *P. vampyrus* and *S. kowalevskii*) because the number of candidates was too large. Whereas function of U1 boxes are diverse (Box A recognizes 5′-splice site in mRNA precursors (Zhuang and Weiner, 1986, and references therein); Box B is part of the U1-70K protein-binding site (Query *et al.*, 1989); Box C is part of the U1-A protein-binding region (Scherly *et al.*, 1989) and the Sm protein-binding region, named 'domain A' (Branlant *et al.*, 1982)), those 5S rRNA boxes are essential for transcription of 5S rRNA itself (Pieler *et al.*, 1987; Bogenhagen, 1993; Hall, 2005). A full color version of this figure is available at the *Heredity* journal online.

instantly into the expected structure as depicted in Figure 1 (observed for all candidates manually), we used constraint folding RNAfold -C. The constraints used are all individually displayed at the Supplemental Page. Alternatively, we created alignments of the previously reported sequences, given in Figure 1 using clustalw (Larkin *et al.*, 2007) and RNAalifold (Hofacker, 2007). Other candidates were marked with additional penalty for further analysis.

Manual inspection. For each organism, alignments were manually examined for irregularities, such as insertions/deletions, indicating a non-functionality.

Final genes and grouping. Finally, each candidate received a classification: satisfying all our filtering criteria indicated that functionality of 5S rRNA sequences was highly likely (Table 1, A-type). If sequence or structure contained slight variations (single point mutations that affected the secondary structure only slightly), then the candidate was declared as B-type.

If both sequence and structure showed several variations, compared with the rest of sequences of that species (for example, indels of at least 5 nt), then the candidate was defined as questionable (Table 1, Q-type), referring to possible pseudogenes.

Even more divergent candidates were deleted from our data sets and not further considered. In fact, we considered these genes to be not functional, to be pseudogenes.

We used the scoring step as a measure of the trustability of each candidate. All sequences, regardless of their score (A, B or Q), were considered in subsequent analyses.

For each organism, fasta, gff and stockholm alignment files are provided on the Supplemental Page.

Orthologous and paralogous 5S rRNA genes

For each taxon, we manually divided the stockholm alignment files into subgroups, defined here as 'blocks' (Table 1).

For the identification of orthologous and paralogous 5S rRNA genes, we used consensus sequences of the blocks and analyzed them with the NeighborNet algorithm (Bryant and Moulton, 2004) and uncorrected *p*-distances in SplitsTree4 (Huson, 1998).

NTS analysis of clusters

Most 5S rRNA occur within 3000 nt. However, to detect a possible correlation of more distantly located 5S rRNAs, we defined 5S rRNA genes being part of one cluster, if and only if they were located on the same chromosome (scaffold or contig) within 10 000 nt independently of their orientation. We wrote postscript files for each taxon to display the genome-wide arrangement. PDF files are provided in the supplement.

The NTS regions <500 nt between two 5S rRNA candidates were aligned with clustalw. Fasta, gff and alignment files are available at the Supplemental Page.

Regulator analysis of 5S rRNA genes

Upstream promoter analysis. We selected the region comprising positions -35 to -25, upstream the 5S rRNA gene (Hallenberg and Frederiksen, 2001; Vizoso *et al.*, 2011) and citations therein, and searched for conserved motifs with MEME (Bailey *et al.*, 2009). We used parameters $-$minw 5 $-$maxw 8 to target a TATA box already described in the literature (see section below). The shuffling of sequences was performed with shuffle -0. Detailed results can be viewed on the Supplemental Page.

Internal promoter analysis. For each stockholm alignment, we created consensus sequences: (A) the most frequent nucleotide was represented in the consensus sequence, and (B) each nucleotide with a frequency $>10\%$ was part of the consensus sequence, following the IUPAC coding system.

Terminator analysis. For the terminator analysis, we analyzed 50 nt downstream of each 5S rRNA candidate and checked with rnabob descriptor the first occurrence of the pattern TTTT. Additionally, conserved motifs were identified by MEME (Bailey *et al.*, 2009) (parameters: $-$minw 6 $-$maxw 20) within the 30 nt downstream. We used only unique sequences per species. Species with >80 different copies were neglected in this analysis, because of complexity reasons.

Linkage between 5S rRNA and other ncRNAs

We downloaded all known ncRNA classes from RFAM (Gardner *et al.*, 2011), and in case of U1 snRNA and 5S rRNA, we included previous literature as mentioned above and searched them in the metazoan genomes with blast (Altschul *et al.*, 1990). Additionally, we scanned the genomes for tRNAs with tRNAscan-SE (Lowe and Eddy, 1997).

As there is, to our knowledge, no established statistical model describing linkage between ncRNAs in a variety of species, we used a simple Gaussian mixture with a variable number of components.

Blast hits are not weighted, that is, hits with an *E*-value below a threshold of 10^{-4} are included, hits above the threshold are excluded. If genomic duplications due to possible assembly artefacts occur, the naive weight given to such a region could point towards linkage, where no real duplication was present. Therefore, we filtered the data: if exactly the identical number of nucleotides were observed between two linked genes, we assumed assembly artefacts (for example, multiple sequenced contigs) and used only one copy.

For each ncRNA gene copy, we test for linkage with 5S, we build a Gaussian mixture ($w_m > 0, \sum w_m = 1$):

$$P(x) = \sum_{m=1}^{k} w_m * \mathcal{N}(\mu_m, \sigma_m).$$

Each Gaussian in the mixture describes the distance μ between a 5S rRNA gene copy and the other gene copy, while σ is the s.d. in this distance. As it is possible that either one 5S gene copy is linked with multiple copies of the other gene, or that multiple pairs of linked 5S rRNAs/other genes exist, we require a *k*-component mixture.

The number of components *k* is determined by increasing *k* from 1 up until no significant improvement in fit is possible. To prevent overfitting, a maximum of 10 Gaussians is allowed, less if the number of data points is lower than 40.

The parameter vector $(\mu_1,\ldots,\mu_k)$, $(\sigma_1,\ldots,\sigma_k)$ is fitted using expectation maximization (Hastie *et al.*, 2001).

RESULTS AND DISCUSSION

For the first time, we present here a complete overview of 5S rDNA in metazoans, including secondary structure prediction, genomic organization, sequence characteristics, putative regulatory motifs and linkage to other ncRNAs. Furthermore, we also found striking features in available mammalian genomes described below. Although this analysis shows many facts that depend on current genome assemblies, the reader should keep in mind that the assemblies of different organisms are extremely variable in terms of completeness and therefore are, at least for the number of copies, hardly comparable. Currently available metazoan genome assemblies very often lack multi-copy regions such as centromeres, telomeres and rRNA operons (Copeland *et al.*, 2009; Dalloul *et al.*, 2010; Alkan *et al.*, 2011). Additionally, two identical gene copies located multiple times in the genome are often merged, or even completely removed (Marz *et al.*, 2008; Alkan *et al.*, 2011). According to Alkan *et al.* (2011), assemblies are in general 16.2% shorter than the reference genome, and 99.1% of validated duplicated sequences are missing from the assembled genome. However, in some assemblies we can find repeated sequences of the same locus, because at the contig or scaffold levels, some genomic regions are covered multiple times. In our analysis, we take these facts into account, and show—as a side effect—how much information we can obtain from genomic sequences when working with multiple-copy genes, regardless of genome assemblies. Available cytogenetic mapping data support our analysis as described in detail below.

Table 1 Number of identified 5S rRNAs

Spec.		No. of Cop.	No. of A	No. of B	No. of Q	No. of Diff.	No. of Blocks	No. of Clus.
hsa	1.1f	18	18	0	0	3	1	1
ppy	1.1f	6	5	0	1	4	1	0
mac	1.1f	12	11	0	1	4	1	0
cjc	1.2c	10	6	0	4	7	3	2
tsy	1.6s	15	5	7	3	15	10	0
oga	1.0s	20	8	8	4	17	4	1
mmu	1.2f	42	41	0	1	6	1	0
rno	1.2f	13	6	7	0	13	5	1
dor	2.0s	27	27	0	0	6	2	0
str	1.1s	5	5	0	0	2	1	0
cpo	1.4s	67	57	8	2	53	3	0
opr	0.6s	7	6	1	0	3	2	2
ocu	0.9s	26	25	1	0	9	1	3
tbe	0.9s	8	4	4	0	6	5	1
fca	0.7s	36	30	6	0	18	2	3
cfa	1.2f	10	8	0	2	9	1	1
vpa	1.1s	8	7	1	0	5	3	0
ttr	1.2s	52	52	0	0	15	1	6
bta	1.2f	10	8	0	2	8	1	0
ssc	1.4f	5	3	2	0	5	5	0
eca	1.3f	5	3	2	0	5	5	0
mlu	0.8s	30	24	2	4	18	3	3
pva	1.2s	28	17	0	11	20	1	8
eeu	1.1s	16	16	0	0	5	1	1
sar	1.0s	17	7	9	1	16	10	0
laf	1.0s	45	24	17	4	36	2	1
ete	1.3s	12	10	1	1	12	5	0
pca	1.3s	10	9	1	0	4	2	1
dno	1.1s	7	7	0	0	3	2	0
cho	1.7s	22	5	9	8	17	4	1
mdo	1.3f	18	17	0	1	9	2	0
meu	1.4s	19	2	16	1	19	5	1
oan	1.4f	23	20	3	0	8	3	0
tgu	1.0f	17	16	0	1	16	2	1
gga	1.1f	6	5	1	0	4	3	1
xtr	2.6s	60	48	10	2	41	9	14
tni	1.1f	54	43	6	5	47	3	16
tru	1.0f	42	41	1	0	23	2	7
gac	10.3s	240	2	0	238	93	4	15
ola	1.2f	3	3	0	0	3	2	0
dre	1.2f	3180	3135	0	45	241	2	43
cmi	2.5c	38	36	1	1	17	3	2
pma	2.0c	344	194	125	25	224	1	74
bfl	1.1s	14	14	0	0	3	1	3
cin	1.1f	48	38	4	6	23	2	12
csa	1.1c	272	236	22	14	133	2	7
odi	1.0s	66	1	0	65	24	1	11
sko	1.0s	1560	80	0	1480	1166	16	223
dme	1.0f	215	98	0	117	24	2	44
dsi	1.0s	14	12	0	2	3	1	4
dse	1.0s	28	16	0	12	9	1	5
der	1.0f	40	31	0	9	8	2	4
dya	1.0s	23	21	0	2	4	1	4
dan	0.9s	64	61	1	2	14	1	13
dps	1.1s	50	50	0	0	5	1	4
dpe	1.0s	60	46	7	6	21	2	18
dwi	0.9s	11	11	0	0	4	1	1
dvi	1.8s	69	68	0	1	6	2	12
dmo	0.9s	61	43	0	18	12	2	6
dgr	1.2s	53	53	0	0	9	1	7

Table 1 (Continued)

Spec.		No. of Cop.	No. of A	No. of B	No. of Q	No. of Diff.	No. of Blocks	No. of Clus.
aga	1.1s	11	10	1	0	3	3	2
aae	0.7c	143	138	3	2	23	1	3
bmo	1.1s	64	58	0	6	39	3	10
tca	1.0c	199	197	1	1	19	1	5
ame	1.0s	53	48	0	5	24	5	10
nvi	0.9f	34	30	4	0	17	3	3
phu	1.0c	31	22	0	9	14	3	6
api	0.7s	50	44	1	5	28	2	3
dpu	1.5s	114	16	1	97	67	4	26
isc	1.3s	7	2	5	0	7	3	1
cre	0.7c	31	0	0	31	6	1	5
cbr	1.1f	7	0	0	7	1	1	2
cbe	0.5c	24	0	0	24	7	1	5
cel	0.9f	13	0	0	13	2	1	1
cja	0.6c	9	0	0	9	2	1	3
hco	0.2c	212	0	0	212	38	3	22
acn	2.2c	10	0	0	10	1	1	3
ppa	1.0c	65	0	0	65	6	1	12
min	0.6c	26	0	0	26	9	1	3
mha	0.8c	12	4	1	7	11	2	2
bma	1.1c	189	0	0	189	19	1	13
tsp	0.1c	424	1	0	423	97	8	0
sma	0.7s	32	22	5	5	23	7	4
sja	0.7c	11	11	0	0	4	1	0
sme	0.8c	66	63	2	1	15	1	13
hro	0.1c	863	732	0	131	264	1	57
cca	0.1c	1584	1493	0	91	410	1	20
apo	2.5c	4	4	0	0	4	1	0
lgi	1.2s	186	166	3	17	88	4	31
aca	2.6s	11	10	1	0	6	1	1
bgl	4.6c	17	16	0	1	7	1	0
esc	78c	5	3	0	2	5	5	0
apa	25c	36	0	0	36	22	3	1
ami	20c	49	0	0	49	41	2	0
nve	0.6s	708	625	7	76	345	3	68
rsp	0.1c	177	162	10	5	55	2	10
tad	0.4s	8	8	0	0	7	3	2

Abbreviatons: aae, *Aedes aegypti*; aca, *Aplysia californica*; acn, *Ancylostoma caninum*; aga, *A. gambiae*; ame, *A. mellifera*; ami, *A. millepora*; apa, *A. palmata*; api, *Acyrthosiphon pisum*; apo, *Alvinella pompejana*; bfl, *Branchiostoma floridae*; bgl, *Biomphalaria glabrata*; bma, *Brugia malayi*; bmo, *Bombyx mori*; bta, *B. taurus*; Blo., number of blocks (groups within an alignment; similar 5S rRNA copies built one block); cbe, *Caenorhabditis brenneri*; cbr, *C. briggsae*; cca, *Capitella sp*; cel, *C. elegans*; cfa, *Canis familiaris*; cho, *Choloepus hoffmanni*; cin, *C. intestinalis*; cja, *C. japonica*; cjc, *Callithrix jacchus Marmoset*; Clus., number of clusters of at least two 5S rRNA-coding regions within 10 000 nt; cmi, *Callorhinchus mili*; cpo, *Cavia porcellus*; cre, *C. remanei*; csa, *C. savignyi*; dan, *Drosophila ananassae*; der, *D. erecta*; dgr, *D. grimshawi*; dme, *D. melanogaster*; dmo, *D. mojavensis*; dno, *Dasypus novemcinctus*; dor, *D. ordii*; dpe, *D. persimilis*; dps, *D. pseudoobscura*; dpu, *D. pulex*; dre, *D. rerio*; dse, *D. sechellia*; dsi, *D. simulans*; dvi, *D. virilis*; Diff., number of different sequences; dwi, *D. willistoni*; dya, *D. yakuba*; eca, *E. caballus*; eeu, *Erinaceus europaeus*; esc, *Euprymna scolopes*; ete, *Echinops telfairi*; fca, *Felis catus*; gac, *Gasterosteus aculeatus*; gga, *G. gallus*; hco, *Haemonchus contortus*; hro, *H. robusta*; hsa, *H. sapiens*; isc, *Ixodes scapularis*; laf, *L. africana*; lgi, *L. gigantea*; mac, *M. mulatta*; mdo, *Monodelphis domestica*; meu, *M. eugenii*; mha, *Meloidogyne hapla*; min, *M. incognita*; mlu, *Myotis lucifugus*; mmu, *Mus musculus*; nve, *Nematostella vectensis*; nvi, *Nasonia vitripennis*; oan, *Ornithorhynchus anatinus*; ocu, *Oryctolagus cuniculus*; odi, *O. dioica*; oga, *Otolemur garnettii*; ola, *O. latipes*; opr, *O. princeps*; pca, *P. capensis*; phu, *P. humanus*; pma, *P. marinus*; ppa, *Pristionchus pacificus*; ppy, *P. pygmaeus*; pva, *P. vampyrus*; rno, *R. norvegicus*; rsp, *Reniera sp*; sar, *Sorex araneus*; sja, *Schistosoma japonicum*; sko, *S. kowalevskii*; sma, *S. mansoni*; sme, *Schmidtea mediterranea*; Spec., Species; spu, *Strongylocentrotus purpuratus*; ssc, *Sus scrofa*; str, *Spermophilus tridecemlineatus*; tad, *T. adhaerens*. tbe, *Tupaia belangeri*; tca, *Tribolium castaneum*; tgu, *Taeniopygia guttata*; tni, *Tetraodon nigroviridis*; tru, *Takifugu rubripes*; tsp, *Trichinella spiralis*; tsy, *T. syrichta*; ttr, *T. truncatus*; vpa, *Vicugna pacos*; xtr, *X. tropicalis*.

We distinguish the total number of candidates (Cop.) to be putatively functional (A), containing variations in sequence or structure (B), and questionable owing to variations in box motifs and secondary structure (Q). The second column depicts the state of each genome assembly: the number is calculated by the known genome size (Gregory, 2012) divided by the number of downloadable nucleotides in **f**inished genomes (chromosomal status), **s**caffolds or **c**ontig.

Arrangement of 5S rRNA copies: number and evolutionary relationship

The overall summary of 5S rRNA copies in animals is depicted in Table 1. We discriminated between three different classes: (A) putative functional genes that passed all our filters, (B) those that showed slight variations in sequence or structure, and (Q) those that remained questionable and might even be possible pseudogenes.

Overall, we identified 12 766 5S rRNA sequences in 97 organisms, ranging from three sequences in the ricefish *Oryzias latipes* to 3180 sequences in the zebrafish *Danio rerio*. Both assemblies are in chromosomal stage. In both cases, real genomes are 1.2-fold larger than the assemblies (Gregory, 2012), see Table 1. The genome coverages of *O. latipes* and *D. rerio* is 10.6 × and ∼30 ×, respectively. Owing to the assembly problems mentioned above, we assume the lower boundary for 5S rRNA copies in these fishes to be about 3180. In general, when the coverage of the genome is at least 8 × and the genome is sorted into chromosomes, it can be considered that the listed number of copies (Table 1) is a lower boundary. Cytogenetic mapping of the *Squalius alburnoides* being closely related to *D. rerio* showed several clusters on three chromosomes (Gromicho *et al.*, 2006), in agreement with the 43 clusters on three chromosomes of the zebrafish in our study. Comparison of fish genomes bring in general difficulties due to polyploidy. The cytogenetic mapping of *Gallus gallus* showed one cluster on chromosome 9 (Cabral-de-Mello *et al.*, 2011), which completely agrees with the one cluster we found also on chromosome 9.

The genome sequence of the most basal deuterostome acorn worm *S. kowalevskii* shows 1166 different copies. Protostomes seem to have, in general, a lower number of 5S rRNA copies. Although the genome of the polychaete worm *Capitella capitata* displays 1584 copies, we assume the real minimal number of 5S rRNA copies to be much smaller, because the genome is on contig stage, which is 10 times larger than the expected genome size (Gregory, 2012), see Supplemental Page. We found 410 different copies of 5S rRNA and we set this value as the minimal copy number in *C. capitata*. By cytogenetic mapping, *Dichotomius* have been shown to consist of a very strong characteristic cluster on chromosome 2 (Cabral-de Mello *et al.*, 2010). The only coleoptera investigated in this manuscript (*Tribolium*) is not assembled on chromosomal level; however, it shows also one huge cluster of 151 5S rRNA copies. Previous reports have shown that copy number is very variable among metazoans: 1700–2000 copies (including pseudogenes) in humans (Sorensen and Frederiksen, 1991), 50–100 copies in *Macaca fascicularis* (Jensen and Frederiksen, 2000), 35–41 copies in the chicken (Daniels and Delany, 2003), 24 000–61 000 copies (including pseudogenes) in three amphibians (24 000 in *Xenopus laevis* (Hilder *et al.*, 1983)) and only three copies in *Plasmodium falciparum* (Shippen-Lentz and Vezza, 1988). However, these estimates relied on the method used and on the ability to differentiate among functional and non-functional copies. Our results do not perfectly agree with these examples as we predicted, in general, a lower number of copies (18 in humans, 12 in *M. mulatta*, 6 in the chicken and 60 *Xenopus tropicalis*).

According to sequence and secondary structure features, we identified different 5S rRNA classes in some genomes as described below. Within species, alignments clearly unveiled disjunct sets, hereafter called 'blocks', in 58 species, see Table 1. We aligned the consensus sequences of the 253 blocks retrieved. In the network obtained, we can distinguish four main 5S rRNA groups, see Figure 2—left.

Orthologous 5S rRNA genes. Vertebrate 5S rRNA sequences are clearly evolutionary separated from other metazoan sequences. Interestingly, basal deuterostomes (Hemichordata, Tunicata and Cephalochordata) and nematodes share high sequence similarity, whereas the sequences of other metazoans (Arthropoda, Lophotrochozoa, Cnidaria, Porifera and Placozoa) clustered into a distinct 5S rRNA group.

Paralogous 5S rRNA genes. When comparing consensus sequences of mammalian 5S rRNA blocks (Figure 2—right), we found, in contrast to non-mammalian sequences, a core 5S rRNA set that comprised at least one sequence of each mammalian species. Sequences within this core set were very similar (nearly no mutations), whereas consensus sequences of the other blocks were relatively divergent (some of them might even be non-functional, such as possibly *Loxodonta africana 2*, see Figure 2—left). No grouping or pattern can be observed in the divergent 5S rRNA set. 5S rRNA seems to have undergone two main evolutionary processes: on the one hand, the data suggest that the long-term evolution of the 5S rRNA genes in mammals is characterized by high selection pressure on housekeeping 5S rRNAs (for example, the 5S rRNA core set) and on the other hand gene diversification, which may provide adaptative potential to environmental change. In other words, we may be facing an evolutionary scenario in which strong purifying selection (and perhaps mechanisms involved in concerted evolution) maintains the integrity of housekeeping 5S rRNAs, whereas birth-and-death processes generate variation through duplications.

The distribution of some orthologous 5S sequences (Figure 2) might be explained by horizontal gene transfer of transposable elements similar to SPIN genes (Syvanen, 2012). However, this is, especially for housekeeping genes, under discussion.

5S rDNA clusters and NTS analysis

In order to study 5S rDNA sequences within species, we analyzed copies separated by less than 10 000 nt (that is, within a 'cluster') in more detail. The number of clusters with at least two 5S rRNAs can be viewed in Table 1. The size of clusters depends on the genome and its assembly, and can be hardly compared.

In many species, we found clusters with differences in the length of their NTS. For example, in the honey bee *Apis mellifera* we found seven copies on contig GroupUn.750 with a constant spacer of 249 nt, whereas contig GroupUn.96 had five copies separated by a 711 nt spacer. Similarly, other species showed NTSs of different sizes in the same contig. This agrees with other species, as previously reported (for example, in molluscs (Vierna *et al.*, 2011), arthropods (Perina *et al.*, 2011) and chordates (Gornung *et al.*, 2007)). In this work, we add to this list more chordate, annelid, arthropod, cephalochordate, placozoan, cnidarian and molluscan species. Sequence orientation and distances between 5S rRNA regions can be obtained for each organism on the Supplemental Page. In the following organisms, we have found 5S rRNA copies that displayed different orientations in the chromosome, a fact that is not in agreement with our expectations according to concerted evolution of repeats within the same cluster. In the cases in which distances among repeats were large (for example, in *X. tropicalis*, *Drosophila melanogaster*, *D. virilis*, *D. mojavensis* and *D. willistoni*), it is not unexpected that gene conversion was unable to homogenize the copies within the cluster. However, in other cases, distances between repeats were small (*Petromyzon marinus*, *Pediculus humanus* and *Trichoplax adhaerens*). This would indicate that the inversions are recent or that the unit of homogenization by gene conversion involves both repeats.

To determine the evolution of 5S rDNA in more detail, we cut and aligned the NTS regions <500 nt (alignment available on the

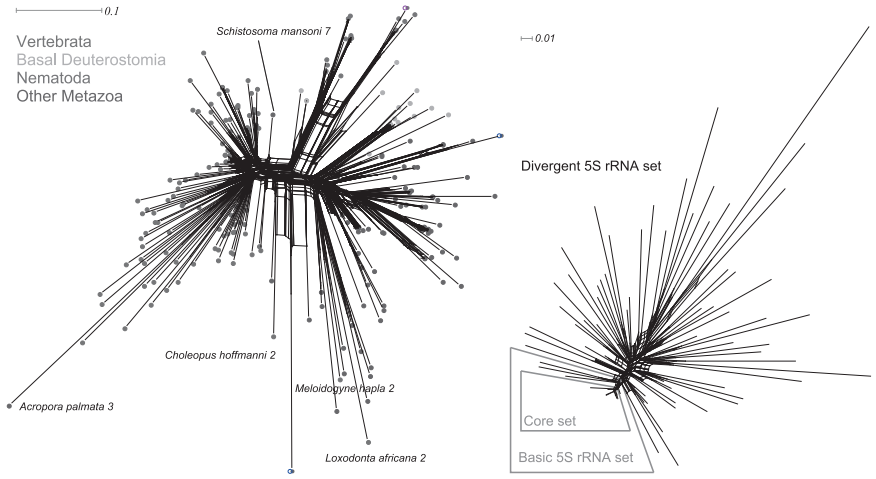


Figure 2 Evolutionary relationships among 5S rRNA. Left: orthologous 5S rRNAs consensus sequences of the 253 5S rRNA blocks (see Table 1). We can distinguish four 5S rRNA groups: vertebrates, basal deuterostomes and nematodes share high sequence similarity, whereas the sequences of other metazoans clustered into a distinct 5S rRNA group. Exceptions are marked with species names and ID of the block. Right: paralogous 5S rRNA genes in mammals. At least one 5S rRNA block consensus sequence of each mammalian genome analyzed is part of the here defined 'basic 5S rRNA set'. The 'core 5S rRNA set' comprises the sequences from the 'basic set' except the one of *Tarsius syrichta*. These basic consensus sequences are highly conserved in sequence. Other 5S rRNA genes are highly divergent. The figure is drawn with the NeighborNet method of the SplitsTree package. All raw data and figures, including taxa names, are downloadable from the Supplemental Page.

Supplemental Page). As hypothezised by Vierna *et al.* (2011), a 5S rDNA sequence that is evolving concertedly within a given cluster can be transposed into another 5S rDNA cluster composed of repeats that are different to that one, but similar among them. After the occurrence of duplications involving both variants, it is possible to obtain an intermixed organization of 5S rDNA, in which NTSs located in the cluster are completely divergent. This is what we report here for some species (*Daphnia pulex* and *D. rerio*, see NTS alignment at Supplemental Page). These findings agree with the widespread idea that 5S rDNA repeats are transposed from one genome location to another (Rooney and Ward, 2005; Datson and Murray, 2006). Intermixed organization of NTS sequences was also found by Gornung *et al.* (2007); Perina *et al.* (2011); Vierna *et al.* (2011) in molluscs, crustacean and fishes species.

Totally unexpectedly, all NTS sequences (divided into four NTS types) retrieved from the mollusc *Lottia gigantea* and from the porifer *Reniera* sp. were almost identical. We blasted these NTSs against various nucleotide databases, but failed to find any similarities with previously reported sequences, such as bacterial/viral insertions. The same picture of very closely related NTS regions is given from the insects *Anopheles gambiae* and *P. humanus*, which is directly in contrast to closely related organisms sharing no related NTS features, such as *Ciona intestinalis* and *C. savignyi* or most of the drosophilids.

We have also retrieved putatively functional and non-functional 5S rRNA sequences within one cluster in many organisms. This has been reported for *D. melanogaster* before (Sharp *et al.*, 1984).

In order to analyze the evolution of the NTS region at the species level, we selected the genus in which the most species were available (*Drosophila*, 12 species). We obtained the following results: (1) NTSs can be divided roughly into 10 different types, according to alignment clustering. In fact, NTS sequences that belong to the same type can be aligned because their degree of divergence is not high; (2) all species display only one type of NTS sequence in their genomes, except *D. mojavensis* and *D. grimshawi*, with two divergent NTS sequence types; (3) the drosophilids do not share their NTS type with their congeners; however, the very recent split species show similar NTS sequences (*D. persimilis*/*D. pseudoobscura* and *D. simulans*/*D. sechellia*); and (4) the different NTS types defined agree with the phylogeny of these 12 species (Drosophila 12 Genomes Consortium *et al.*, 2007). Considering these results and the high degree of conservation of the 5S rRNA copies, we hypothesize an evolutionary scenario in which the long-term evolution of 5S rDNA in the genus *Drosophila* is driven by strong selection over the 5S rRNA copies, gene duplications and transpositions that generate new NTS loci, and homogenizing mechanisms within each array. The divergent NTSs retrieved from *D. mojavensis* and *D. grimshawi* could also point toward the occurrence of ancestral polymorphism. Birth-and-death evolution with a fast gene turnover, concerted evolution and mixed models combined with strong selection can explain these results.

Internal promoter analysis

The internal 5S rRNA boxes (Figure 1) are essential for transcription of 5S rRNA itself (Pieler *et al.*, 1987; Bogenhagen, 1993; Hall, 2005). For instance, the C2H2 zinc finger protein TFIIIA binds *Xenopus* 5S rRNA internal promoters (named Box A, intermediate element and Box C (Pieler *et al.*, 1987)), see Bogenhagen (1993) and Hall (2005). To analyze the occurrence of these essential boxes in our set of 5S rRNAs, we obtained consensus sequences for each block of 5S rRNAs taking into account the IUPAC coding system. In Figure 3a, consensus sequence of these consensus sequences is depicted. We found several conserved regions, which are shown highlighted. In the 5′-region of the 5S rRNA molecule, we found two strikingly conserved motifs: CAUAC (9–14 nt) and GAA (21–23 nt). Furthermore, we detected high conservation in the center of the structure, between positions 39 and 60. Finally, the 3′-end was highly conserved (GUG). By taking into account the secondary structure, the regions interacting with polymerase coincide. Our results confirm previous reports by Pieler *et al.* (1987); Sharp and Garcia (1988); Vizoso *et al.* (2011):

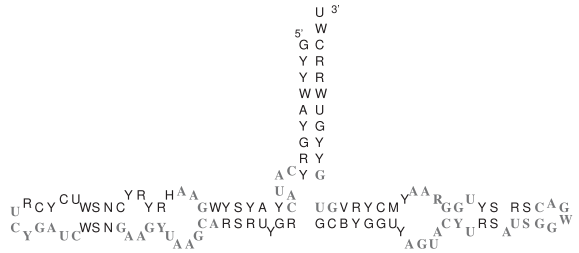


Figure 3 Consensus sequence and structure of metazoan 5S rRNA. Conserved motifs are highlighted. A full color version of this figure is available at the *Heredity* journal online.

the CAUAC covers the ICR-I, which was described for *D. melanogaster* by Pieler *et al.* (1987). The known ICR-II region (GUCCGAUC) is almost identical to our YCRUCYGAUC-motif, which occurs in all so far known 5S rRNAs and shows therewith very high conservation. For the ICR-III (as described in *D. melanogaster*: GAAAUUAAGCAGCG), we detected that the first U is often replaced by C, whereas the second U, which was reported for the molluscs as M (A or C), is highly conserved (Vizoso *et al.*, 2011). In the case of ICR-IV, the known sequence for *D. melanogaster* (Pieler *et al.*, 1987) does only match partially our consensus sequence.

Interestingly, most of our conserved motifs and structure can also be found in land plants, such as the CAUAC motif (Wicke *et al.*, 2011). A comparison of the consensus structure of metazoans and land plants (Wicke *et al.*, 2011) shows that they highly correspond with each other. There are slight differences concerning the length of hairpins, such as smaller internal loops in the 5′-end of the metazoan 5S, instead of two larger loops in plants (referred as B and C in Wicke *et al.* (2011)).

Potential regulatory upstream regions

We analyzed 40 nt upstream of each of our 5S rRNA sequences using MEME, and detected three conserved motifs. The motif WTAAA was retrieved in 35 species (1.7e–172, 370 sequences, see Supplemental Page). The TATA-box TATAAT was found in 13 genomes (5.2e −026, 71 sequences, see Supplemental Page). They were mainly protostomes, but we also found this motif in several sequences of the tunicate *Oikopleura dioica*. One copy per genome with this motif was found in three more deuterostomes. Interestingly, a TATA-like motif located at positions around −30 to −25 from the transcription start site is essential for efficient 5S rRNA transcription *in vitro* in *Caenorhabditis elegans* and *C. briggsae* (Nelson *et al.*, 1998), *Neurospora crassa* (Tyler, 1987) and *D. melanogaster* (Sharp and Garcia, 1988). Upstream motif analysis has also detected TATA-like elements in several bivalve groups, with different degrees of conservation, at positions −30 to −25 (Vizoso *et al.*, 2011), and in fishes (Martins and Wasko, 2004). Finally, we also found the motif TCTTGRGG (5.5e −077, 61 sequences) in 28 species (Supplemental Page). Among them, there were 25 mammals and three other species (the polychaete *Capitella sp.*, the leech *Helobdella robusta* and the trematode *Schistosoma mansoni*). It is known that the mammalian 5S rRNA gene has a 12-nucleotide upstream promoter, the D-box (GGCTCTTGGGGC), which is important for efficient transcription *in vitro* and *in vivo* in humans (Hallenberg and Frederiksen, 2001). The D-box is located in positions −32 to −21 nt in humans and in *M. fascicularis*; in positions −33 to −22 nt in mouse and rat; and in positions −36 to −25 nt in hamster (Hallenberg and Frederiksen, 2001). However, we failed to find this motif in the remaining eight mammalian genomes (*Rattus norvegicus*, *Dipodomys ordii*, *Ochotona princeps*, *Tursiops truncatus*, *Equus caballus*, *Echinops telfari*, *Procavia capensis* and *Macropus eugenii*). Shuffling of all the sequences resulted in GC-rich motifs with large *E*-values (>360).

Terminator analysis

In Bogenhagen and Brown (1981); Huang and Maraia (2001) and Richard and Manley (2009), the transcription termination signal for 5S rDNA is described as one or more TTTT stretches. We analyzed 50 nt downstream of 5S rRNA-coding regions using MEME. We found three common patterns (Figure 4, for more specific patterns, see Supplement).

As expected, we obtained 272 sequences with poly-T tail in 40 species. However, we also found 99 polyadenylated sequences in 26 species. We assume these are 5S rRNA copies that have been transcribed and inserted back into the genome as reported previously for mouse and rat (Drouin, 2000). These 99 sequences are found among all deuterostomes and are quite a large fraction compared with the 272 poly-T sequences. Sequence and structure of these 5S rRNA copies are completely conserved to putatively functional ones, suggesting that (a) these genes are transcribed (as reported for mouse and rat in Drouin (2000) and/or (b) the possible insertion by retroelements was very recent in terms of evolutionary time. For *Drosophila* only, we found 10 nucleotides located between the 5S rRNA-coding region and the poly-T tail. This is in agreement with Sharp and Garcia (1988), demonstrating that 135 nt are transcribed, whereas the mature transcript has only 120 nt.

This leads to the general question of how frequent longer intermediate transcripts are, compared with the size of the mature RNA. We tried to answer this question by plotting the distance from the RNA-coding region to the first TTTT motif, defined as terminator (Bogenhagen and Brown, 1981; Huang and Maraia, 2001; Richard and Manley, 2009). Figure 5 shows that, in general, deuterostomes have longer intermediate transcripts than protostomes. Human and fly share the feature of a 11-nt-longer preprocessed transcript. However, there are exceptions: in *Dipodomys* we found only 1 nt between RNA-coding region and TTTT, and in the most basal metazoan *T. adhaerens*, we found a distance of 7–13 nt.

Martins and Wasko (2004) described for some fish species an additional conserved downstream motif (GAAACAA), but its function is not known. However, we only found this motif in very few of our conserved terminal regions, in unrelated species.

Finally, we detected an interesting feature in *Acropora palmata*, as analysis of terminal sequences, Figure 6, suggested a systematic insertion or deletion of thymines downstream of 5S rRNA-coding regions.

5S rRNA and protein interactions

Scripture and Huber (1995) showed for *X. laevis* the interaction of protein eL5 to 5S rRNA helix III and loop C, as well as a adenine pairing in helix III. With our results, we were able to show that this might be a metazoan-wide feature, as these regions are conserved for all sequences (Figure 3). Interestingly, we were able to show that the interaction of transcription factor IIIA seems not to be consistent for all metazoans. This is in agreement with Lu *et al.* (2003), who showed that TFIIIA is interacting with helix V and helix II. Both regions were not conserved among our examined metazoan sequences. Although many data on 5S rRNA structure and interactions exists, its function is still not clearly elucidated (Smirnov *et al.*, 2008).

Linkage between 5S rRNA and other ncRNAs

For 12 ncRNA genes out of all families of Rfam, we detected significant evolutionary linkage to 5S rDNA. For each of the ncRNAs, we provide a number of mixture model plots on the Supplemental Page.

Genomic linkage of 5S rDNA to major *spliceosomal RNA* genes has been reported for nematodes (Vahidi *et al.*, 1988), crustaceans (U1, U5 (Pelliccia *et al.*, 2001; Marz *et al.*, 2008)), molluscs (U1, U2, (Cross and Rebordinos, 2005; Vierna *et al.*, 2011)) and fishes (U1, U2, U5 (Manchado *et al.*, 2006)). In this work, we detected a weak correlation to U1, U4 and U5 snRNA genes, and a slightly more evolutionary consistent linkage to U2. 5S rDNA was found in many organisms linked to U6 snRNA genes, probably owing to the high repetition of this snRNA, in addition to the copy number of 5S rDNA itself. Although we were able to identify linkages to U1 snRNA genes, a consistent linkage feature between organisms was not detected. To refuse the assumption that U1 snRNA genes were not detected correctly by blast, we analyzed this gene family the same way as 5S rDNAs (Figure 1). However, we obtained the same weak results on linkage.

The splice leader (*SL RNA*) is known to be linked to 5S rDNA in nematodes (Nilsen *et al.*, 1989; Zeng *et al.*, 1990) and protists (Keller *et al.*, 1992). In this work, we found a strong linkage between SL RNA genes and 5S rDNA in all nematodes and platyhelminthes, which has never been reported before. Linkages were found to be sense and antisense (see Supplemental Page).

Furthermore, we were able to detect linkages to many *tRNA* genes, as already described by Freire *et al.* (2010) and Vizoso *et al.* (2011). Linkages to *5.8S rRNA genes, SRP RNA genes, Y RNA genes, Histone3 RNA and 7SK genes* were detected rarely compared with all investigated organisms (see Supplemental Page). However, the linkage of 5S rRNAs and transposable elements, such as SRP RNA genes, seems to appear at least in primates and glires.

Additionally, we found linkages between 5S rDNA and one of eight miRNAs and two snoRNAs. One example is miRNA-562, which has never been described to be linked to 5S rDNA before (Figure 7). In this case, there are nine Gaussians that form the mixture. Evidence for possible linkage can be found in primates, mouse and bat, with most other species showing either no or no consistent information. We show four linkages consistent within primates at distances of around −1648, −1221, −440 and 714 nt. A fifth linkage at a distance of 3000 nt does not seem to be as well-supported by the data. Additionally, the linkages we found between mir-562 and 5S rDNA in mammal species can be traced back to eutherians.

In the bat *P. vampyrus*, the number of 5S rRNAs is constant to evolutionary related organisms (28 copies, see Table 1), but an expansion of mir-562 can be detected. Interestingly, each of the bat 5S rRNAs is linked (with variable distance) to at least one of the mir-562 copies.

Although a long-lasting stable linkage of 5S RNA is known for non-animals (for example, 5S–45S linkage in Bryophytes (Wicke *et al.*, 2011)), we were not able to find a stable linkage within metazoans.

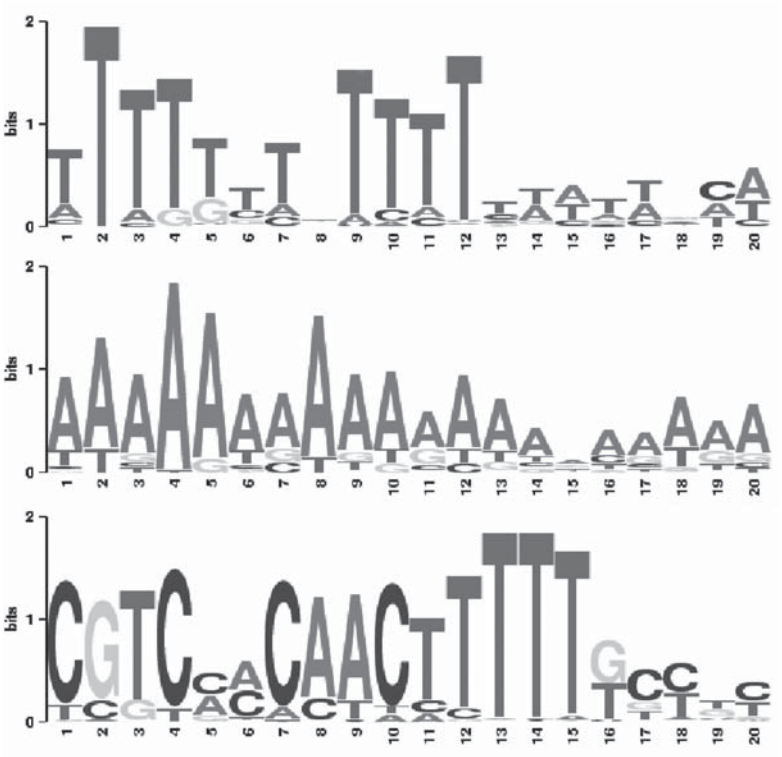


Figure 4 Terminator pattern detected within the 30 nt downstream of 5S rRNA candidates, using MEME. Top: previously known poly-T region in 272 sequences with an *E*-value of 4.8×10^{-168}. Middle: newly detected poly-A region in 99 sequences (1.6×10^{-270}). Bottom: conserved region in 112 insect, one *Sus scrofa* and 17 bat *P. vampyrus* sequences (2.3×10^{-361}) in agreement with Sharp *et al.* (1984) and Sharp and Garcia (1988).

Figure 5 Evolutionary boxplot of distances between 5S rRNA-coding region and first occurrence of the downstream located TTTT motif. TTTT, number of sequences with TTTT within the first 30 nt; sum, number of terminator sequences analyzed; diff, number of different terminator sequences (50 nt analyzed); and uni, number of terminator sequences, which have at least one identical copy.

```
apa5_1   TTTTTTTTTTTTCTTTTTGATTACACCATTGAA
apa4_1   -TTTTTTTTTTTCTTTTTGATTACACCGTTGAA
apa3_3   --TTTTTTTTTTCTTTTTGATTACACCATTGAA
apa2_15  ---TTTTTTTTTCTTTTTGATTACACCATTGAA
apa1_5   ----TTTTTTTTCTTTTTGATTACACCATTGAA
ama1_2   -----TTTTTTTCTTTTTGATTACACCATTGAA
```

Figure 6 *A. palmata* sequences downstream of 5S rRNA-coding regions, and systematic insertion or deletion of thymines.

Unfortunately, the general sparsity (as, for example, in Figure 7) of the data requires manual intervention, which is starting the EM algorithm with multiple different start values for the parameters (μ, σ) or selection between two different proposed mixtures (different number k of peaks).

We are currently working on a more extensive statistical description of linkage between genes with the aim to automate the whole decision process regarding linkage analysis of a set of genomes that should greatly simplify analyses like the present one. Details are described in Marz and Höner zu Siederdissen (2013).

CONCLUSION

For the first time, a comprehensive set of putatively functional 5S rDNA sequences from current metazoan genome assemblies is published. This large amount of data allowed us to study metazoan

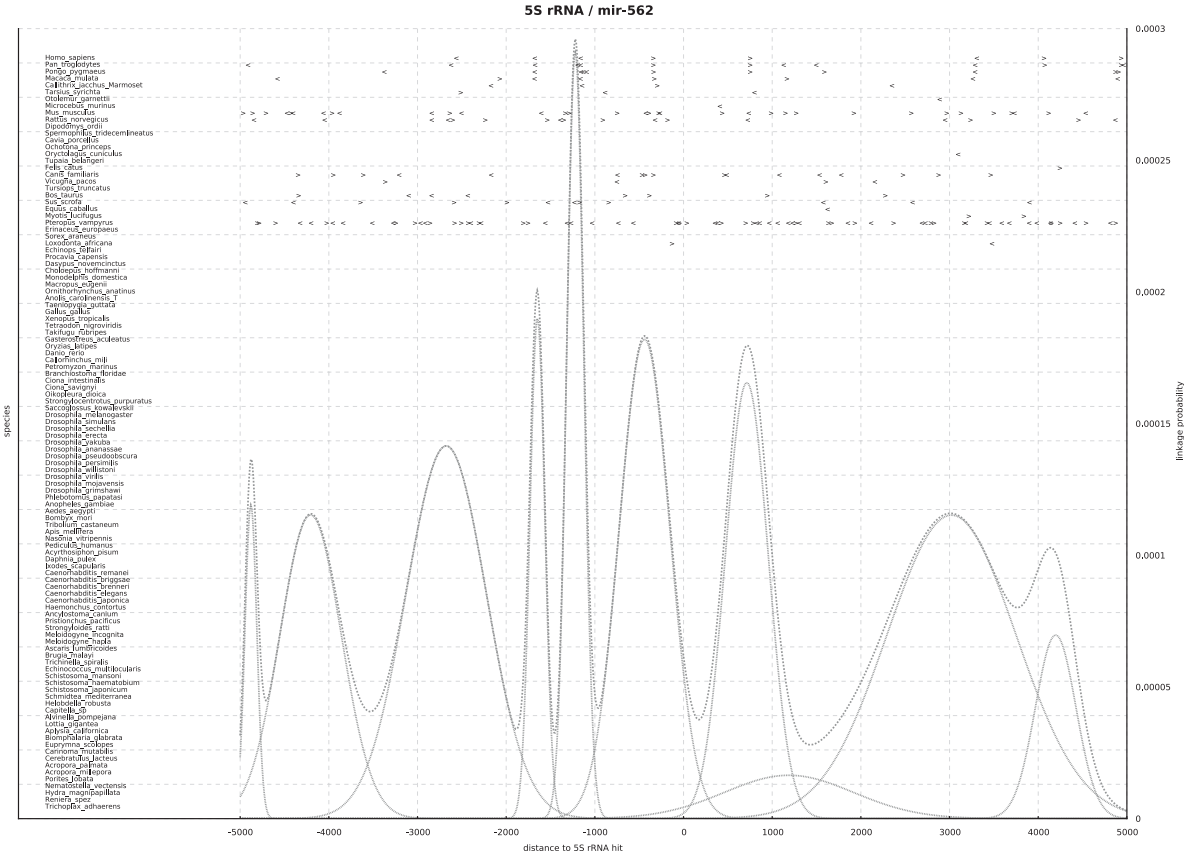


Figure 7 Evolutionary-conserved linkages of 5S rRNA and mir-562. Distance between any annotated mir-562 (denoted as '>' if sense; or '<' if antisense) and any 5S rRNA (adjusted at $x=0$, not displayed) in animal genomes (y axis, left). The y axis (right) denotes the probability (see curve functions) that these ncRNA gene linkages are of evolutionary importance. In primates, mir-562 is located conserved upstream (-1648 nt, antisense; -1221 nt, antisense; -440 nt; sense) and downstream (714 nt, sense) of 5S rRNA. In the bat *P. vampyrus*, the number of 5S rRNAs is constant (28 copies, see Table 1), but an expansion of mir-562 can be detected. Interestingly, each of the bat 5S rRNAs is linked (with variable distance) to at least one of the mir-562 copies.

5S rDNA diversity in great detail, following a systematic approach and from an evolutionary perspective.

Among our main conclusions, we showed that 5S rRNA-coding regions in mammals are divided into two types that we name 'housekeeping 5S rRNAs', which are very conserved, and 'flexible 5S rRNAs', being much more variable. In addition, we found several paralog 5S rRNA-coding sequences in many species (58 out of 97 genomes).

We also reported a flexible genome organization of 5S rDNA, as it was found either (1) in clusters, linked to other ncRNAs, (2) in homogeneous clusters, with similar NTS sequences, (3) in heterogeneous clusters, with divergent NTS sequences, (4) in clusters in which coding regions displayed opposite orientations and (5) as dispersed copies. Interestingly, several species displayed more than one of these features.

The unexpected similarity found among NTS sequences of some distantly related taxa is unclear and intriguing. On the one hand, we might hypothesize that those sequences are unidentified elements that have a molecular role in the cell and therefore have been conserved along evolution. On the other hand, they might be the result of horizontal gene transfer events.

Remarkably, even though we found 5S rDNA to be linked to several ncRNAs in many species, we failed to detect a stable linkage throughout animal evolution.

As the biological meaning of various features that were found to characterize metazoan 5S rDNA are still to be elucidated, this work opens up very interesting possibilities for molecular biology research. For example, the meaning of different promoter and terminator sequences that have been found should be unveiled.

Dimarco *et al.* (2012) have recently reported that different 5S rRNA variants are transcribed in a sea urchin species. In the same way, whether transcription of paralog 5S rRNA sequences reported in the present work occurs equally in all tissues deserves with no doubts to be studied in detail.

Finally, our work supports the conclusions of various recent reports (Rooney and Ward, 2005; Kalendar *et al.*, 2008; Vierna *et al.*, 2009; Freire *et al.*, 2010; Perina *et al.*, 2011; Vierna *et al.*, 2011; Vizoso *et al.*, 2011), in which it was shown that the evolutionary patterns of 5S rDNA in animals are complex and cannot be explained only in the light of concerted evolution. Birth-and-death processes, selection, homogenizing mechanisms typically involved in concerted evolution

and horizontal gene transfer events seem to be responsible of the diversity of this multigene family in metazoans.

DATA ARCHIVING

All the data can be accessed via the following link: http://www.rna.uni-jena.de/supplements/5SRNA/index.html.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

JV was funded by a 'María Barbeito' fellowship and a travel grant from Xunta de Galicia and Universidade da Coruña (Spain). MM was funded by the Carl-Zeiss-Stiftung. This work was supported in part by DFG-Graduiertenkolleg 1384 'Enzymes and multienzyme complexes acting on nucleic acids', DFG project MA-5082/1 (SW and MM) and by the Austrian FWF, project 'SFB F43 RNA regulation of the transcriptome' (CHzS).

Alkan C, Sajjadian S., Eichler EE (2011). Limitations of next generation genome sequence assembly. *Nat Methods* **8**: 61–65.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L *et al.* (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: 202–208.

Barzotti R, Pelliccia F, Rocchi A (2003). Identification and characterization of U1 small nuclear RNA genes from two crustacean isopod species. *Chromosome Res* **11**: 365–373.

Bogenhagen DF (1993). Proteolytic footprinting of transcription factor TFIIIA reveals different tightly binding sites for 5S RNA and 5S DNA. *Mol Cell Biol* **13**: 5149–5158.

Bogenhagen DF, Brown DD (1981). Nucleotide sequences in Xenopus 5S DNA required for transcription termination. *Cell* **24**: 261–270.

Branlant C, Krol A, Ebel JP, Lazar E, Haendler B, Jacob M (1982). U2 RNA shares a structural domain with U1, U4, and U5 RNAs. *EMBO J* **1**: 1259–1265.

Bryant D, Moulton V (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**: 255–265.

Cabral-de Mello DC, Moura RC, Martins C (2010). Chromosomal mapping of repetitive DNAs in the beetle Dichotomius geminates provides the first evidence for an association of 5S rRNA and histone H3 genes in insects, and repetitive DNA similarity between the B chromosome and A complement. *Heredity (Edinb)* **104**: 393–400.

Cabral-de-Mello DC, Moura RC, Martins C (2011). Cytogenetic mapping of rRNAs and histone H3 genes in 14 species of Dichotomius (Coleoptera, Scarabaeidae, Scarabaeinae) beetles. *Cytogenet Genome Res* **134**: 127–135.

Cohen S, Agmon N, Sobol O, Segal D (2010). Extrachromoso mal circles of satellite repeats and 5S ribosomal DNA in human cells. *Mobile DNA* **1**: 1–11.

Copeland CS, Marz M, Rose D, Hertel J, Brindley PJ, Santana CB *et al.* (2009). Homology-based annotation of non-coding RNAs in the genomes of *Schistosoma mansoni* and *Schistosoma japonicum*. *BMC Genomics* **10**: 464–464.

Cross I, Rebordinos L (2005). 5S rDNA and U2 snRNA are linked in the genome of *Crassostrea angulata* and *Crassostrea gigas* oysters: does the (CT)n(GA)n microsatellite stabilize this novel linkage of large tandem arrays? *Genome* **48**: 1116–1119.

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le Ann *et al.* (2010). Multi-platform next-generation sequencing 698 of the domestic turkey (Meleagris gal lopavo): genome assembly and analysis. *PLoS Biol* **8**: e1000475.

Daniels LM, Delany ME (2003). Molecular and cytogenetic organization of the 5S ribosomal DNA array in chicken (*Gallus gallus*). *Chromosome Res* **11**: 305–317.

Datson PM, Murray BG (2006). Ribosomal DNA locus evolution in Nemesia: transposition rather than structural rearrangement as the key mechanism? *Chromosome Res* **14**: 845–857.

Dimarco E, Cascone E, Bellavia D, Caradonna F (2012). Functional variants of 5S rRNA in the ribosomes of common sea urchin Paracentrotus lividus. *Gene* **508**: 21–25.

Drosophila 12 Genomes ConsortiumClark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA *et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

Drouin G (2000). Expressed retrotransposed 5S rRNA genes in the mouse and rat genomes. *Genome* **43**: 213–215.

Drouin G, Moniz de Sá M (1995). The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol Biol Evol* **12**: 481–493.

Drouin G, Tsang C (2012). 5S rRNA gene arrangements in protists: a case of nonadaptive evolution. *J Mol Evol* **74**: 342–351.

Eickbush TH, Eickbush DG (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**: 477–485.

Eirin-Lopez JM, Fernanda Ruiz M, Gonzalez-Tizon AM, Martinez A, Sanchez L, Mendez J (2004). Molecular evolutionary characterization of the mussel Mytilus histone multigene family: first record of a tandemly repeated unit of five histone genes containing an H1 subtype with "orphon" features. *J Mol Evol* **58**: 131–144.

Freire R, Arias A, Insua AM, Méndez J, Eiŕin-López JM (2010). Evolutionary dynamics of the 5S rDNA gene family in the mussel Mytilus: mixed effects of birth-and-death and concerted evolution. *J Mol Evol* **70**: 413–426.

Fujiwara M, Inafuku J, Takeda A, Watanabe A, Fujiwara A, Kohno S *et al.* (2009). Molecular organization of 5S rDNA in bitterlings (Cyprinidae). *Genetica* **135**: 355–365.

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S *et al.* (2011). Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res* **39**: 141–145.

Gillespie JJ, Johnston JS, Cannone JJ, Gutell RR (2006). Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of *Apis mellifera* (Insecta: Hymenoptera): structure, organization, and retrotransposable elements. *Insect Mol Biol* **15**: 657–686.

Gongadze GM (2011). 5S rRNA and ribosome. *Biochemistry* **76**: 1450–1464.

Gornung E, Colangelo P, Annesi F (2007). 5S ribosomal RNA genes in six species of Mediterranean grey mullets: genomic organization and phylogenetic inference. *Genome* **50**: 787–795.

Gregory T (2012). Animal Genome Size Database. (http://www.genomesize.com/)

Griffiths-Jones S (2005). RALEE–RNA ALignment editor in Emacs. *Bioinformatics* **21**: 257–259.

Gromicho M, Coutanceau JP, Ozouf-Costaz C, Collares-Pereira MJ (2006). Contrast between extensive variation of 28S rDNA and stability of 5S rDNA and telomeric repeats in the diploid-polyploid *Squalius alburnoides* complex and in its maternal ancestor Squalius pyrenaicus (Teleostei, Cyprinidae). *Chromosome Res* **14**: 297–306.

Hastie T, Tibshirani R, Friedman J (eds) (2001). *The Elements of Statistical Learning — Data Mining, Inference, and Prediction*, 2nd edn Springer.

Hall TM (2005). Multiple modes of RNA recognition by zinc finger proteins. *Curr Opin Struc Biol* **15**: 367–373.

Hallenberg C, Frederiksen S (2001). Effect of mutations in the upstream promoter on the transcription of human 5S rRNA genes. *Biochim Biophys Acta* **1520**: 169–173.

Hallenberg C, Frederiksen S (2001). Effect of mutations in the up stream promoter on the transcription of human 5S rRNA genes. *Biochim Biophys Acta* **1520**: 169–173.

Hilder VA, Dawson GA, Vlad MT (1983). Ribosomal 5S genes in relation to C-value in amphibians. *Nucleic Acids Res* **11**: 2381–2390.

Hofacker IL (2003). Vienna RNA secondary structure server. *Nucleic Acids Res* **31**: 3429–3431.

Hofacker IL (2007). RNA consensus structure prediction with RNAalifold. *Methods Mol Biol* **395**: 527–544.

Huang Y, Maraia RJ (2001). Comparison of the RNA polymerase III transcription machinery in *Schizosaccharomyces pombe*, Saccha769 romyces cerevisiae and human. *Nucleic Acids Res* **29**: 2675–2690.

Huson DH (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**: 68–73.

Jensen LR, Frederiksen S (2000). The 5S rRNA genes in *Macaca fascicularis* are organized in two large tandem repeats. *Biochim Bio phys Acta* **1492**: 537–542.

Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O *et al.* (2008). Cassandra retrotransposons carry indepen dently transcribed 5S RNA. *Proc Natl Acad Sci USA* **105**: 5833–5838.

Keller M, Tessier LH, Chan RL, Weil JH, Imbault P (1992). In Euglena, spliced-leader RNA (SL-RNA) and 5S rRNA genes are tandemly repeated. *Nucleic Acids Res* **20**: 1711–1715.

Komiya H, Kawakami M, Takemura S (1981). Nucleotide sequence of 5S ribosomal RNA from the posterior silk glands of Bombyx mori. *J Biochem* **89**: 717–722.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Layat E, Sáez-Vásquez J, Tourmente S (2012). Regulation of Pol I-transcribed 45S rDNA and Pol III-transcribed 5S rDNA in Arabidopsis. *Plant Cell Physiol* **53**: 267–276.

Little RD, Braaten BC (1989). Genomic organization of human 5S rDNA and sequence of one tandem repeat. *Genomics* **4**: 376–383.

Lowe TM, Eddy SR (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

Lu D, Searles MA, Klug A (2003). Crystal structure of a zinc finger-RNA complex reveals two modes of molecular recognition. *Nature* **426**: 96–100.

Manchado M, Zuasti E, Cross I, Merlo A, Infante C, Rebordinos L (2006). Molecular characterization and chromosomal mapping of the 5S rRNA gene in Solea senegalensis: a new linkage to the U1, U2, and U5 small nuclear RNA genes. *Genome* **49**: 79–86.

Martins C, Wasko AP (2004). *Organization and Evolution of 5S Ribosomal DNA in the Fish Genome*. Nova Science Publishers, Inc.

Marz M, Höner zu Siederdissen C (2013). *Statistical approach for evolutionary gene linkage detection*. (in preparation).

Marz M, Kirsten T, Stadler PF (2008). Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol* **67**: 594–607.

Nelson DW, Linning RM, Davison PJ, Honda BM (1998). 5′-flanking sequences required for efficient transcription in vitro of 5{S} {RNA} genes, in the related nematodes *Caenorhabditis elegans) and Caenorhabditis briggsae*. *Gene* **218**: 9–16.

Nei M, Rooney AP (2005). Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**: 121–152.

Nilsen TW, Shambaugh J, Denker J, Chubb G, Faser C, Putnam L *et al.* (1989). Characterization and expression of a spliced leader RNA in the parasitic nematode Ascaris lumbricoides var. suum. *Mol Cell Biol* **9**: 3543–3547.

Pelliccia F, Barzotti R, Bucciarelli E, Rocchi A (2001). 5S ribosomal and U1 small nuclear RNA genes: a new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences. *Genome* **44**: 331–335.

Perina A, Seoane D, González-Tizón AM, Rodŕiguez-Fariña F, Marťinez-Lage A (2011). Molecular organization and phylogenetic analysis of 5S rDNA in crustaceans of the genus *Pollicipes* reveal birth- and death evolution and strong purifying selection. *BMC Evol Biol* **11**: 304–304.

Pieler T, Hamm J, Roeder RG (1987). The 5S gene internal control region is composed of three distinct sequence elements, organized as two functional domains with variable spacing. *Cell* **48**: 91–100.

Pieler T, Hamm J, Roeder RG (1987). The 5S gene internal control region is composed of three distinct sequence elements, organized as two functional domains with variable spacing. *Cell* **48**: 91–100.

Query CC, Bentley RC, Keene JD (1989). A specific 31-nucleotide domain of U1 RNA directly interacts with the 70K small nuclear ribonucleoprotein component. *Mol Cell Biol* **9**: 4872–4881.

Richard P, Manley JL (2009). Transcription termination by nuclear RNA polymerases. *Gene Dev* **23**: 1247–1269.

Rooney AP, Ward TJ (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. *Proc Natl Acad Sci USA* **102**: 5084–5089.

Scherly D, Boelens W, van Venrooij WJ, Dathan NA, Hamm J, Mattaj IW (1989). Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. *EMBO J* **8**: 4163–4170.

Schnare MN, Gray MW (1999). A candidate U1 small nuclear RNA for trypanosomatid protozoa. *J Biol Chem* **274**: 23691–23694.

Scripture JB, Huber PW (1995). Analysis of the binding of Xenopus ribosomal protein L5 to oocyte 5 S rRNA. The major determinants of recognition are located in helix III-loop C. *J Biol Chem* **270**: 27358–27365.

Shambaugh JD, Hannon GE, Nilsen TW (1994). The spliceosomal U small nuclear RNAs of Ascaris lumbricoides. *Mol Biochem Parasit* **64**: 349–352.

Sharp S, Garcia A, Cooley L, Söll D (1984). Transcriptionally active and inactive gene repeats within the *D. melanogaster* 5S RNA gene cluster. *Nucleic Acids Res* **20**: 7617–7632.

Sharp SJ, Garcia AD (1988). Transcription of the *Drosophila melanogaster* 5S RNA gene requires an upstream promoter and four intragenic sequence elements. *Mol Cell Biol* **8**: 1266–1274.

Sharp SJ, Garcia AD (1988). Transcription of the *Drosophila melanogaster* 5S RNA gene requires an upstream promoter and four intragenic sequence elements. *Mol Cell Biol* **8**: 1266–1274.

Shippen-Lentz DE, Vezza AC (1988). The three 5S rRNA genes from the human malaria parasite *Plasmodium falciparum* are linked. *Mol Biochem Parasit* **27**: 263–273.

Smirnov AV, Entelis NS, Krasheninnikov IA, Martin R, Tarassov IA (2008). Specific features of 5S rRNA structure - its interactions with macromolecules and possible functions. *Biochemistry* **73**: 1418–1437.

Sorensen PD, Frederiksen S (1991). Characterization of human 5S rRNA genes. *Nucleic Acids Res* **19**: 4147–4151.

Syvanen M (2012). Evolutionary implications of horizontal gene transfer. *Annu Rev Genet* **46**: 341–358.

Thomas J, Lea K, Zucker-Aprison E, Blumenthal T (1990). The spliceosomal snRNAs of *Caenorhabditis elegans*. *Nucleic Acids Res* **18**: 2633–2642.

Tyler BM (1987). Transcription of *Neurospora crassa* 5 S rRNA genes requires a TATA box and three internal elements. *J Mol Biol* **196**: 801–811.

Vahidi H, Curran J, Nelson DW, Webster JM, McClure MA, Honda BM (1988). Unusual sequences, homologous to 5S RNA, in ribosomal DNA repeats of the nematode Meloidogyne arenaria. *J Mol Evol* **27**: 222–227.

Vierna J, González-Tizón AM, Marťinez-Lage A (2009). Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death and selection in *Ensis* razor shells (Mollusca: Bivalvia). *Biochem Genet* **47**: 635–644.

Vierna J, Jensen KT, Marťinez-Lage A, González-Tizón AM (2011). The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae). *Heredity* **107**: 127–142.

Vizoso M, Vierna J, González-Tizón AM, Marťinez-Lage A (2011). The 5S rDNA gene family in mollusks: characterization of transcriptional regulatory regions, prediction of secondary structures, and long-term evolution, with special attention to Mytilidae mussels. *J Hered* **102**: 433–447.

Wicke S, Costa A, Muñoz J, Quandt D (2011). Restless 5S: the re-arrangement(s) and evolution of the nuclear ribosomal DNA in land plants. *Mol Phylogenet Evol* **61**: 321–332.

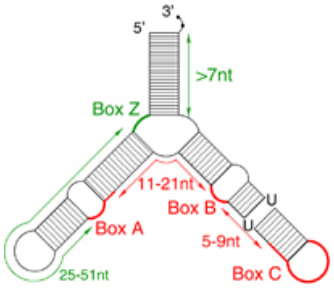
Zeng WL, Alarcon CM, Donelson JE (1990). Many transcribed regions of the Onchocerca volvulus genome contain the spliced leader sequence of *Caenorhabditis elegans*. *Mol Cell Biol* **10**: 2765–2773.

Zhuang Y, Weiner AM (1986). A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* **46**: 827–835.

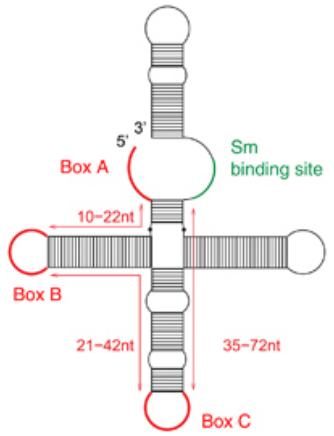
Úbeda-Manzanaro M., Merlo M. A., Palazón J. L., Sarasquete C., Rebordinos L. (2010). Sequence characterization and phylogenetic analysis of the 5S ribosomal DNA in species of the family Batrachoididae. *Genome* **53**: 723–730.

Ørum H, Nielsen H, Engberg J (1991). Spliceosomal small nuclear RNAs of Tetrahymena thermophila and some possible snRNA-snRNA base-pairing interactions. *J Mol Biol* **222**: 219–232.

Figure 1



Organisms	Box A	Box B	Box C	Box Z	Lit.
Consensus	YAAGCA	UUAGUA	GGGWGA		
Vertebrates	U.....	.....	...A..		Martins and Wasko (2004)
Xenopus sp.	C.....	.....	...A..		Pieler *et al.* (1987)
Bombyx mori	C.....	.C....	...U..		Komiya *et al.* (1981)
Honey bee	C....U	.....	...U..		Gillespie *et al.* (2006)
Drosophila sp.	U.....	.....	...G..		Sharp and Garcia (1988)
Razor Shells	U.....	.N....	...U..		Vierna *et al.* (2011)
Plasmodium	U.....	C.....	...A..		Shippen-Lentz and Vezza (1988)
Search motif	-AAGC-	--AGUA	GGGNGA	NNNNNCAUA	
Mutations	1	1	1	1	



Organisms	Box A	Box B	Box C	Sm bind. site	Lit.
Consensus	AWACUUACCUG	GAUCA	AUUGCAC	UAAUUUNUG	
Vertebrates	.U.....	.....	.....	.....G..	Marz *et al.* (2008)
"Basal Deut."	.U.....	.....	.....	.....C..	Marz *et al.* (2008)
Insects	----.....	.....	.....	.....U..	Marz *et al.* (2008)
A. aquaticus	.U.....	.....	.....	.....C..	Barzotti *et al.* (2003)
"worms"	-A.....	.....	.....	.....U..	Marz *et al.* (2008)
C. elegans	.A.....	.....	.....	.....U..	Thomas *et al.* (1990)
A. lumbricoides	.A.....	.....	.....	.....U..	Shambaugh *et al.* (1994)
Trypanosoma	.U.....	.....	.....	.....G..	Schnare and Gray (1999)
Razor Shells	.U.....	.....	.....	C.....A..	Vierna *et al.* (2011)
Tetrahymena	--.....	.....	.....	.....C..	Ørum *et al.* (1991)
Search motif	----UUACCUG	GAUCA	AUUGCAC	-AAUUUNU-	
Mutations	2	2	2	1	

Figure 2 left

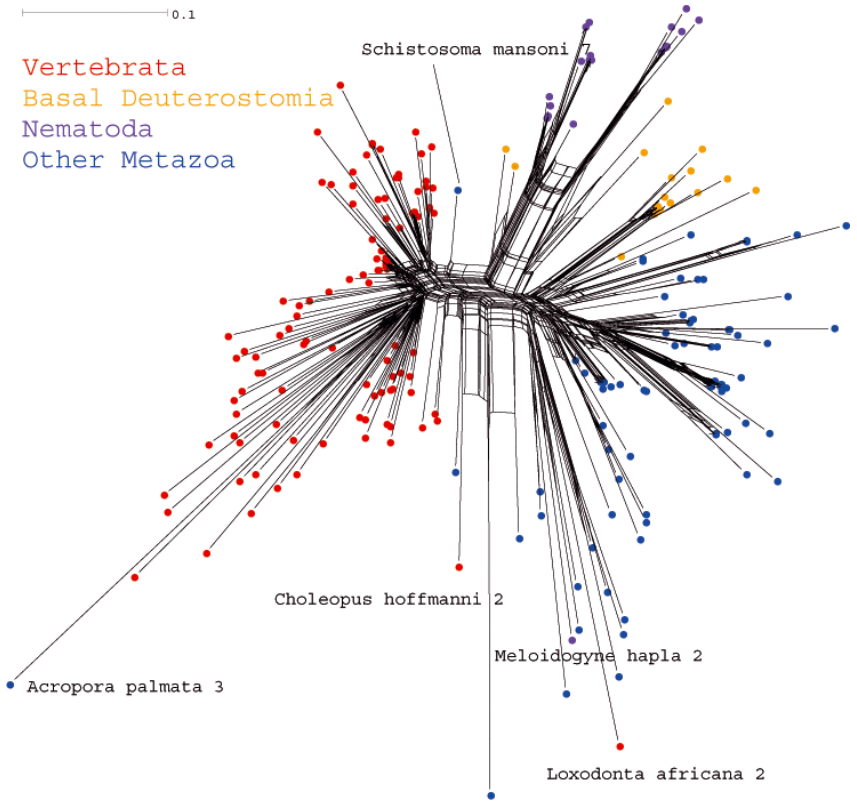


Figure 3

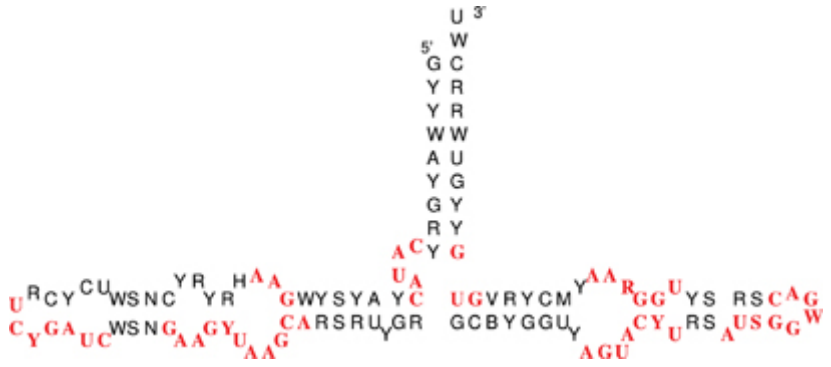


Figure 4

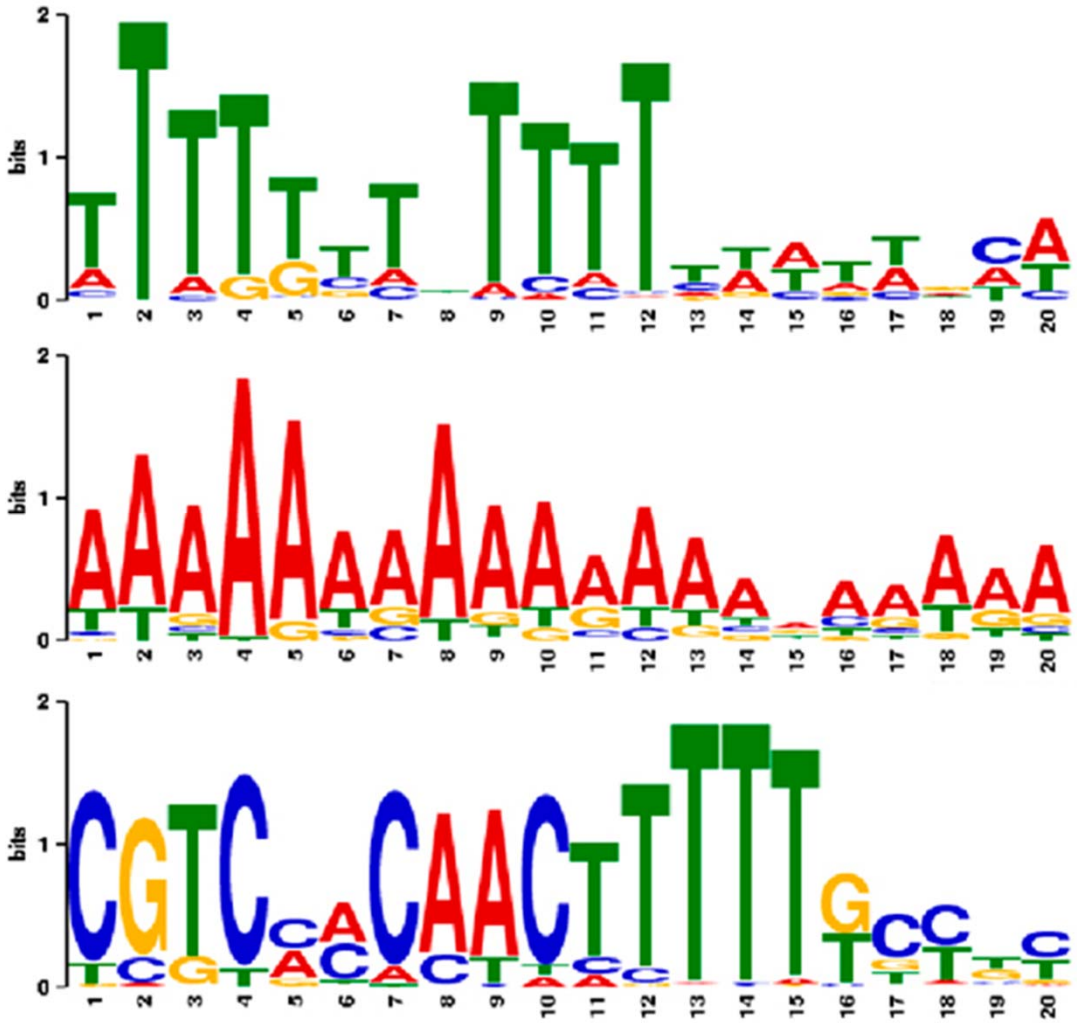


Figure 6

```
apa5_1    TTTTTTTTTTTTCTTTTTGATTACACCATTGAA
apa4_1    -TTTTTTTTTTTCTTTTTGATTACACCGTTGAA
apa3_3    --TTTTTTTTTTCTTTTTGATTACACCATTGAA
apa2_15   ---TTTTTTTTCTTTTTGATTACACCATTGAA
apa1_5    ----TTTTTTTCTTTTTGATTACACCATTGAA
ama1_2    -----TTTTTTCTTTTTGATTACACCATTGAA
```

4.2 Cytogenetic characterisation of the razor shells *Ensis directus* (Conrad, 1843) and *E. minor* (Chenu, 1843) (Mollusca: Bivalvia)

Ana M. González-Tizón, Verónica Rojo, Joaquín Vierna, K. Thomas Jensen, Emilie Egea, Andrés Martínez-Lage (2013) Cytogenetic characterisation of the razor shells *Ensis directus* (Conrad, 1843) and *E. minor* (Chenu, 1843) (Mollusca: Bivalvia). Helgoland Marine Research 67:73-82.

Bibliometrics 2012 JCR Science Edition

Impact factor: 1.444

Marine & Freshwater Biology: Q3

Oceanography: Q3

Helgol Mar Res (2013) 67:73–82
DOI 10.1007/s10152-012-0305-4

ORIGINAL ARTICLE

Cytogenetic characterisation of the razor shells *Ensis directus* (Conrad, 1843) and *E. minor* (Chenu, 1843) (Mollusca: Bivalvia)

Ana M. González-Tizón · Verónica Rojo · Joaquín Vierna · K. Thomas Jensen · Emilie Egea · Andrés Martínez-Lage

Received: 19 October 2011 / Revised: 12 January 2012 / Accepted: 10 April 2012 / Published online: 26 April 2012

Abstract The European razor shell *Ensis minor* (Chenu 1843) and the American *E. directus* (Conrad 1843) have a diploid chromosome number of 38 and remarkable differences in their karyotypes: *E. minor* has four metacentric, one metacentric–submetacentric, five submetacentric, one subtelocentric and eight telocentric chromosome pairs, whereas *E. directus* has three metacentric, two metacentric–submetacentric, six submetacentric, six subtelocentric and two telocentric pairs. Fluorescent in situ hybridisation (FISH) using a major ribosomal DNA probe located the major ribosomal genes on one submetacentric chromosome pair in both species; FISH with a 5S ribosomal DNA (5S rDNA) probe rendered one chromosomal (weak) signal for *E. minor* and no signal for *E. directus*, supporting a more dispersed organisation of 5S rDNA compared to the major ribosomal genes. The vertebrate telomeric sequence $(TTAGGG)_n$ was located on both ends of each chromosome, and no interstitial signals were detected. In this work, a comparative karyological analysis was also performed between the four *Ensis* species analysed revealing that the three European species studied so far, namely *E. minor*, *E. siliqua* (Linné 1758) and *E. magnus* Schumacher 1817 show more similarities among them than compared to the American species *E. directus*. In addition, clear karyotype differences were found between the morphologically similar species *E. minor* and *E. siliqua*.

Keywords Razor shells · Karyotype · FISH · 18S-5.8S-28S rDNA · 5S rDNA · Telomeric sequence

Communicated by Heinz-Dieter Franke.

A. M. González-Tizón (✉) · V. Rojo · J. Vierna · A. Martínez-Lage
Department of Cell and Molecular Biology, Universidade da Coruña, A Fraga 10, 15008 La Coruña, Spain
e-mail: hakuna@udc.es

K. T. Jensen
Marine Ecology, Department of Bioscience, Aarhus University, Ole Worms Allé 1, Building 1135, 8000 Aarhus C, Denmark

E. Egea
Centre d'Océanologie de Marseille, Station Marine d'Endoume, Chemin de la Batterie des Lions, 13007 Marseille, France

Introduction

The genus *Ensis* Schumacher 1817 (Mollusca: Bivalvia: Pharidae) is composed of about 12 extant species that live on fine sand, silt or mud bottoms off the European, African and American coasts. In Europe, four species are native, *E. ensis* (Linné 1768), *E. magnus* Schumacher 1817 [syn. *E. arcuatus* (Jeffreys 1865)], *E. minor* (Chenu 1843) and *E. siliqua* (Linné 1758) and one introduced, *E. directus* (Conrad 1843) [syn. *E. americanus* (Gould 1870)]. This species was introduced to the German Bight at the end of the 1970 s from Atlantic North America, probably through ballast water (Cosel et al. 1982). In European coastal waters, the distribution areas of the different *Ensis* species are mostly overlapping and there are few areas occupied by only one species (Cosel 2009) though *E. directus* prefers brackish waters. The European *E. minor* has often been confused with its homonym *E. minor* Dall 1899 (which is native to the SE United States), although they constitute different species. Additional taxonomic confusion is due to the fact that both European *E. minor* and *E. siliqua* are very similar in terms of shell morphology. Cosel (2009) states "*E. minor* was frequently treated under the name *E. siliqua* or as a subspecies of that taxon; however, along the Atlantic coast the two species occur sympatrically with

only very few possible intergrades which occasionally were found at the south coast of Brittany (Pl. 8f–k) and also a few at Ile de Ré, Charente Maritime. They are looking superficially like *E. minor* but mostly have the rounded posterior cross-section of *E. siliqua* (Pl. 8e). Only molecular research will elucidate this".

The two species studied in this work are distributed as follows: *E. minor* from British North Sea coast from the east coast of Scotland to southern England, The Netherlands, and the Channel; European Atlantic coast from the NW part of Wales southward to North Morocco and throughout the Mediterranean. The introduced *E. directus* is now well established in Europe and occurs from Denmark to France, England and Sweden (Cosel 2009).

Whereas recent molecular analyses have been performed in a number of *Ensis* (Varela et al. 2007, 2009; Vierna et al. 2009, 2010, 2011), there is only one report about karyotypes of these species (specifically only in *E. magnus* and *E. siliqua* by Fernández-Tajes et al. 2008). It is worthwhile to mention that there are more than 20,000 different molluscan species worldwide distributed in aquatic habitats, and only about 200 species have been cytogenetically studied (i.e. karyotyped). This is due to the fact that cytogenetic studies of molluscs are usually complicated because the difficulties derived of the very low mitotic index in adult tissues and the problems to gain higher mitotic indexes. Recent articles (since year 2000) dealing about cytogenetic characterisation of commercial marine molluscs have been performed in different species of mussels (Vitturi et al. 2000; Martínez-Lage et al. 2002; Iqbal et al. 2008; Pérez-García et al. 2010, 2011), clams (González-Tizón et al. 2000; Martínez et al. 2002; Plohl et al. 2002; Hurtado and Pasantes 2005; Leitao et al. 2006; Wang and Guo 2007, 2008), razor shells (Fernández-Tajes et al. 2003, 2008), oysters (Xu et al. 2001; Cross et al. 2005; Wang et al. 2005a, b; Huang et al. 2007a, b; Zhang et al. 2007), cockles (Leitao et al. 2006), and pectinids (Pauls and Afonso 2000: López-Piñón et al. 2005; Odierna et al. 2006; Huang et al. 2007a, b; Zhang et al. 2008).

Cytogenetic analyses are important as they provide information about the number and morphology of chromosomes, the differential distribution of euchromatin-heterochromatin regions, the occurrence of chromosomal re-arrangements along evolution, phylogenetic relatedness between taxa, etc., and they help to clarify species status, which is extremely important in conservation biology. Therefore, they can be applicable in aquaculture as cytogenetic techniques, mainly FISH, are a significant part of genomic research to facilitate the construction of linkage maps, which are useful to identify loci of interest in economic marine species.

In the present study, we describe the chromosome number and morphology, and the location of the major ribosomal loci (18S-ITS1-5.8S-ITS2-28S) and the telomeric sequences in *E. minor* and *E. directus*. The location of 5S ribosomal DNA (5S rDNA) in *E. minor* is also provided. In addition, we perform a comparative karyological analysis among *E. minor*, *E. directus*, *E. magnus* and *E. siliqua* in order to infer phylogenetic relationships based on karyotype differences and to clarify *E. minor*–*E. siliqua* taxonomic status.

Materials and methods

Biological material

Specimens of *E. minor* were collected from La Capte (43°02′N, 6°09′E) and Bandol (43°08′N, 5°46′E) (both localities on the Eastern Gulf of Lion, France), and those of *E. directus* were caught in Vester Vedsted (55°17′N, 8°38′E) (Denmark). In the laboratory, animals were fed with a suspension of *Isochrysis galbana* and *Tetraselmis* sp. microalgae for 10–15 days. Specimens were identified according to Cosel (2009).

Chromosome preparation

Metaphases were obtained from gill tissue of adult specimens after treatment with colchicine solution (0.005 %) for 6–8 h. Gills were dissected and treated twice with 0.56 % KCl solution for 15 min. After fixation in ethanol–glacial acetic acid (3:1), cells were dissociated in 45 % acetic acid and dropped onto slides heated at 43 °C. Metaphases were stained with 4 % Giemsa in phosphate buffer pH 6.8 and photographed using a Nikon Microphot-FXA microscope equipped with a NIS-Elements D 3.10 software and a digital camera DS-Qi1Mc.

Karyotyping

Chromosome measurements were performed using the Leica Chantal image analysis software system described in González-Tizón et al. (2000). Chromosome measurements were carried out in 12 metaphases from 12 individuals in both species. Mean value of the length of the chromosome arms and the mean value for their total chromosome lengths were calculated for each of the chromosome pairs. The relative length (100 × chromosome length/total haploid length), the centromeric index (100 × length of short arm/total chromosome length), the mean value and the standard error (standard deviation/number of individuals)$^{1/2}$ of the relative lengths and centromeric indices were also calculated. Karyotypes were arranged by decreasing size and classified according to the centromeric index, following the nomenclature of Levan et al. (1964).

Fluorescent in situ hybridisation (FISH)

Chromosomal location of rDNA loci was performed by FISH as described in González-Tizón et al. (2000). The DNA probe pDM 238 from *Drosophila melanogaster* (Roiha et al. 1981), containing the repeat unit 18S-5.8S-28S rDNA, was labelled by nick translation with digoxigenin-11-dUTP (Roche) for chromosome mapping of major ribosomal genes. The 5S rDNA probe was obtained by PCR using the primers 5S-Univ-F (5′-ACCGGTGTTTTC AACGTGAT) and 5S-Univ-R (5′-CGTCCGATCACCG AAGTTAA) designed by Vierna et al. (2009). These primers had opposite orientations and were separated by 3 bp. The probe was labelled by PCR with digoxigenin-11-dUTP (Roche).

Telomeric FISH was carried out as in Plohl et al. (2002) with the $(TTAGGG)_{22}$ probe labelled with digoxigenin by a standard PCR procedure. Slides were counterstained with propidium iodide (50 ng/mL antifade) and visualised and photographed using a Nikon Microphot-FXA microscope equipped with a NIS-Elements D 3.10 software and a digital camera DS-Qi1Mc.

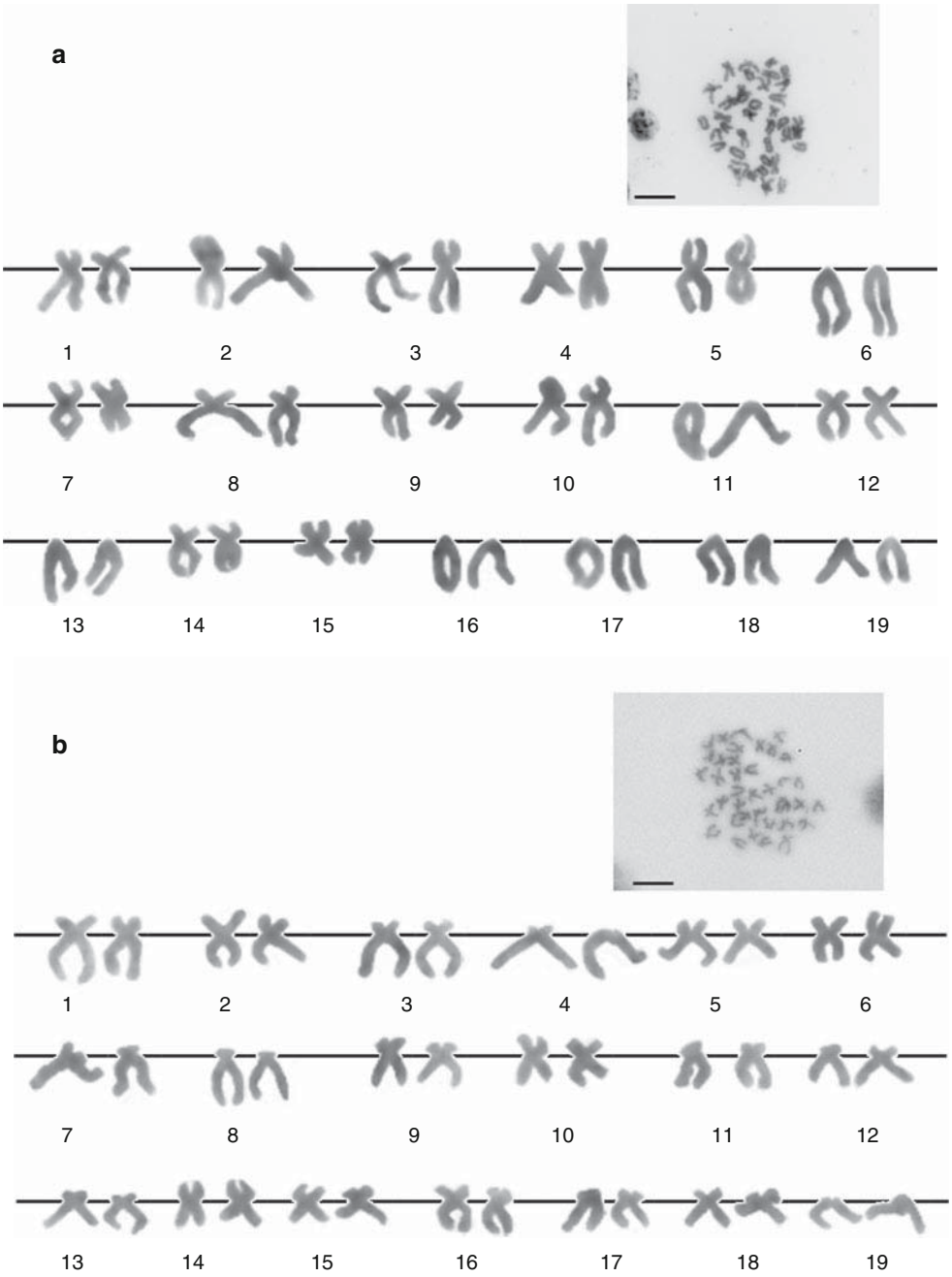


Fig. 1 Karyotypes of **a** *Ensis minor* and **b** *E. directus* (both $2n = 38$ chromosomes). *Bar* 10 μm

Results

Analysis of 38 metaphases from 15 specimens of *E. minor* and 42 metaphases from 12 specimens of *E. directus* revealed a diploid chromosome number for each species of $2n = 38$ (Fig. 1a, b). For karyotyping, ten well-spread metaphases were paired on the basis of chromosome size and centromere position and used for chromosome measurements and classification. In *E. minor*, relative length varied from 3.40 to 6.99, and in *E. directus*, it ranged from 3.46 to 6.93 (Table 1). The karyotype of *E. minor* consisted of four metacentric chromosome pairs, one metacentric–submetacentric, five submetacentric, one subtelocentric and eight telocentric (Figs. 1a, 3a). The species *E. directus* had three metacentric, two metacentric–submetacentric, six submetacentric, six subtelocentric and two telocentric chromosome pairs (Figs. 1b, 3b).

In both species, FISH using a 18S-ITS1-5.8S-ITS2-28S probe revealed the location of a major ribosomal locus on the short arm of one submetacentric chromosome pair at an interstitial position (Figs. 2a, b, 3a, b). Telomeric signals appeared at the end of all chromosomes on both *Ensis* species (Fig. 2c, d), and a signal for the 5S rDNA gene family was located at a subtelomeric region of one medium-size telocentric chromosome pair in *E. minor* (Fig. 2e), whereas *E. directus* did not show fluorescent signal after 5S rDNA hybridisation. In our previous studies (Vierna et al. 2009, 2011), PCR amplifications of 5S rDNA in *E. directus* generated multiple fragments, with sizes ranging between 406 and 739 bp, corresponding to 5S rDNA variants that differed in the length of the nontranscribed spacer region (NTS). Additional PCR products were identified as dimer and trimer sequences formed by two and three contiguous monomers, respectively. The species *E. minor* was not studied in these reports, but differences in 5S rDNA organisation are not to be expected among *E. minor* and the other European species analysed (i.e. *E. ensis*, *E. siliqua* and *E. magnus*).

Discussion

This study reveals that *E. minor* and *E. directus* have a diploid chromosome number of 38 chromosomes, which is coincident with those previously reported on their congenerics *E. magnus* and *E. siliqua* (Fernández-Tajes et al. 2008), and with the majority of the karyotypes studied within the Heterodonta bivalves. However, there are some exceptions as *Kidderia minuta* (Thiriot-Quievreux et al. 1988a), *Lasaea australis* (Thiriot-Quievreux 1992), *Spaherium*

Table 1 Chromosome measurements and classification

Ensis minor				*Ensis directus*			
	RL	CI	Class		RL	CI	Class
1	6.99 ± 0.06	32.73 ± 0.61	sm	1	6.93 ± 0.03	26.18 ± 0.79	sm
2	6.61 ± 0.03	35.52 ± 0.76	sm	2	6.56 ± 0.07	41.09 ± 0.37	m
3	6.36 ± 0.03	33.74 ± 0.26	sm	3	6.24 ± 0.06	29.43 ± 0.41	sm
4	6.24 ± 0.04	31.36 ± 0.52	sm	4	6.16 ± 0.04	22.53 ± 0.57	st
5	6.05 ± 0.03	39.88 ± 0.56	m	5	5.99 ± 0.11	33.80 ± 0.60	sm
6	5.88 ± 0.08	0.11 ± 0.01	t	6	5.74 ± 0.05	37.85 ± 0.37	m–sm
7	5.75 ± 0.03	37.95 ± 0.65	m–sm	7	5.66 ± 0.07	33.28 ± 0.79	sm
8	5.39 ± 0.05	0.12 ± 0.01	t	8	5.60 ± 0.04	8.16 ± 0.35	t
9	5.29 ± 0.06	39.61 ± 0.63	m	9	5.31 ± 0.03	22.58 ± 0.84	st
10	5.22 ± 0.05	33.46 ± 0.51	sm	10	5.25 ± 0.06	38.84 ± 0.30	m
11	5.11 ± 0.05	0.12 ± 0.01	t	11	5.17 ± 0.04	22.56 ± 0.47	st
12	4.87 ± 0.08	17.75 ± 1.48	st	12	5.02 ± 0.04	21.66 ± 0.71	st
13	4.84 ± 0.04	2.65 ± 0.84	t	13	4.87 ± 0.04	38.06 ± 0.72	m–sm
14	4.78 ± 0.07	38.53 ± 0.58	m	14	4.66 ± 0.06	38.56 ± 0.58	m
15	4.60 ± 0.08	39.98 ± 0.62	m	15	4.55 ± 0.05	30.12 ± 1.24	sm
16	4.53 ± 0.03	0.14 ± 0.01	t	16	4.41 ± 0.04	23.33 ± 0.96	st
17	4.23 ± 0.04	0.15 ± 0.01	t	17	4.22 ± 0.03	14.38 ± 0.95	st
18	3.87 ± 0.04	0.16 ± 0.01	t	18	4.20 ± 0.04	27.23 ± 0.66	sm
19	3.40 ± 0.05	1.46 ± 0.41	t	19	3.46 ± 0.06	7.67 ± 0.34	t

RL relative length, *CI* centromeric index, *Class* classification, *m* metacentric, *sm* submetacentric, *st* subtelocentric, *t* telocentric

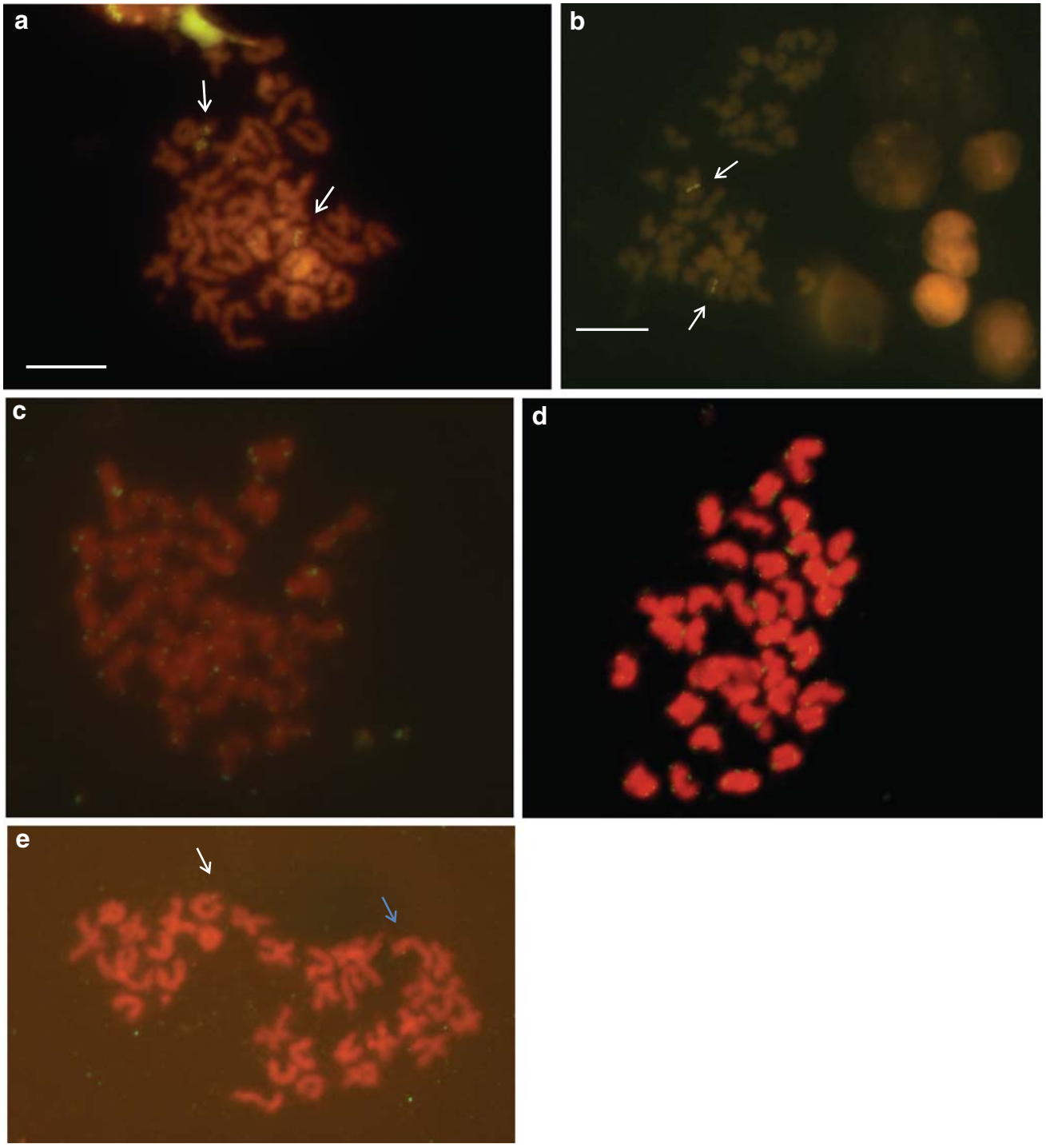


Fig. 2 Metaphases of *Ensis minor* after **a** FISH with an 18S-ITS1-5.8S-ITS2-28S ribosomal DNA probe, **c** FISH with a telomeric probe, **e** FISH with a 5S ribosomal DNA probe. Metaphases of *E. directus* after **b** FISH with an 18S-ITS1-5.8S-ITS2-28S ribosomal DNA probe, **d** FISH with a telomeric probe. *Bar* 10 μm

corneum (Petkeviciute et al. 2006) and *Cyclinia sinensis* (Wang et al. 2001), all of them with $2n = 36$ chromosomes, *Cyclocardia astartoides* (Thiriot-Quievreux et al. 1990) with $2n = 30$, and some *Lasaea* species with different levels of polyploidy (Thiriot-Quievreux et al. 1988b, 1989; Ó Foighil and Thiriot-Quievreux 1999).

The karyotypes of *E. minor* and *E. directus* have telocentric pairs, which is not very usual in Heterodonta.

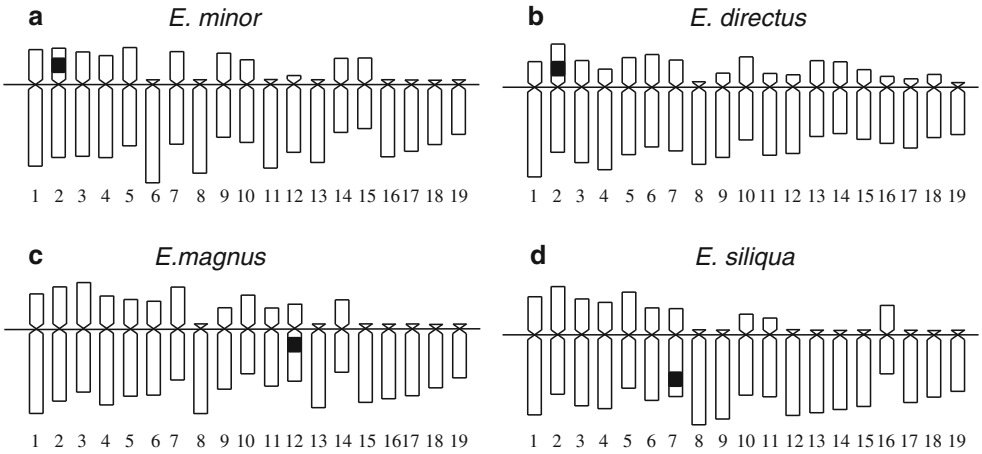


Fig. 3 Idiograms of **a** *Ensis minor*, **b** *E. directus*, **c** *E. magnus* and **d** *E. siliqua*. *Black refill* represents the location of the 18S-ITS1-5.8S-ITS2-28S ribosomal loci. **c**, **d** Data taken from Fernández-Tajes et al. (2008)

To our knowledge, only the karyotypes of 29 Heterodonta species reported so far belonging to 12 different families have telocentric chromosomes (Table 2). The number of telocentric pairs varies from one in *Donax trunculus* (Martínez et al. 2002), *Lasaea colmani* (Ó Foighil and Thiriot-Quievreux 1999), *Scrobicularia plana* (Cornet and Soulard 1989), *Sinonovacula constricta* (Wang et al. 1998), and *Ruditapes aureus* (= *Venerupis aurea*) (Borsa and Thiriot-Quievreux 1990; Carrilho et al. 2011) to 19 (all the complement) in *Mulinia lateralis* (Wang and Guo 2008).

The comparative analysis of *Ensis* karyotypes shows that the diploid chromosome number is the same for the four species and that the number of telocentrics is higher in the European ones than in the American species (nine pairs in *E. siliqua,* eight in *E. minor* and seven in *E. magnus*, whereas *E. directus* has only two telocentric pairs). In the four species, the number of metacentric and submetacentric pairs is similar: 11 in *E. directus* and *E. magnus* and 10 in *E. minor* and *E. siliqua*, being the major differences in the number of subtelocentrics: one pair in *E. minor* and *E. magnus* and six pairs in *E. directus*, whereas *E. siliqua* has no subtelocentrics. These differences lead to karyotypes with numbers of chromosome arms of: 72 in *E. directus*, 62 in *E. magnus*, 60 in *E. minor* and 58 in *E. siliqua*, this representing a major divide between the American species (*E. directus*) and the European species (*E. magnus*, *E. siliqua* and *E. minor*). According to White (1978) who pointed out that karyotypes with higher proportion of metacentrics are generally considered as more primitive (plesiomorphic) and show relative more chromosome stability than karyotypes with few metacentric chromosomes (apomorphic), the results obtained for the *Ensis* species suggest that *E. directus* is the species with the most ancestral karyotype, whereas the karyotypes of the native European species may be more recent. Differences in the number of bi-armed and mono-armed chromosomes lead us to think that structural re-arrangements (involving loss of chromosomal arms and emergence of telocentrics, fusion of telocentric chromosomes and reciprocal or robertsonian translocations, or inversions) could have occurred during the evolution of *Ensis* species, as suggested Wang and Guo (2004) for Pectinidae species and Wang and Guo (2008) for *M. lateralis*.

These results confirm *E. siliqua* and *E. minor* as separate species, clarifying their taxonomic status, and may be very useful in future programmes on aquaculture and conservation of these species.

Concerning the major ribosomal loci, these have been mapped using FISH in 10 Heterodonta species: *Cerastoderma edule* (Insua et al. 1999), *Donax trunculus* (Martínez et al. 2002), *Solen marginatus* (Fernández-Tajes et al. 2003), *Dosinia exoleta* (Hurtado and Pasantes 2005), *Mercenaria mercenaria* (Wang and Guo 2007), *E. magnus* and *E. siliqua* (Fernández-Tajes et al. 2008) (Fig. 3c, d), *M. lateralis* (Wang and Guo 2008) and *Tapes rhomboides* and *V. aurea* (Carrilho et al. 2011). All of them showed one major ribosomal locus, except *S. marginatus* (Fernández-Tajes et al. 2003) and *M. lateralis* (Wang and Guo 2008), which had two loci. The four *Ensis* species showed only one chromosomal interstitial signal on one submetacentric pair (Fig. 3). In *E. minor* and *E. directus*, the fluorescent signal appeared interstitially located on the p arm of chromosome number 2, whereas in *E. magnus* and *E. siliqua*, the signal is on q arm of chromosome number 12 and 7, respectively. These differences in location of major ribosomal genes could be explained by translocations, as suggested Wang and Guo (2004) for pectinids. In *E. minor* and *E. directus*, FISH signals were stronger on one of the homologous chromosomes than on the other, which, as pointed out by Xu et al. (2001) is probably due to random variation in FISH or the differences (loss or gain) in the number of rDNA repeats.

The FISH experiments using 5S rDNA probes were only obtained until now in three species belonging to the subclass Heterodonta, the cockle *C. edule* (Insua et al. 1999), which revealed nine hybridisation signals, and the clams

Table 2 List of Heterodonta species with telocentric chromosomes

Taxa an diploid chromosome number	No. of telocentric pairs	Authors	No. of 18S–28S loci (FISH)
Family CARDIIDAE			
Cerastoderma edule ($2n = 38$)	3	Insua and Thiriot-Quievreux (1992), Insua et al. (1999)	1
Family CARDITIDAE			
Cyclocardia astartoides ($2n = 30$)	10	Thiriot-Quievreux et al. (1990)	
Family CYAMIIDAE			
Kidderia bisulcata ($2n = 38$)	5	Thiriot-Quievreux et al. (1988a)	
Kidderia minuta ($2n = 36$)	5	Thiriot-Quievreux et al. (1988a)	
Family DONACIDAE			
Donax trunculus ($2n = 38$)	1	Martínez et al. (2002)	1
Family LASAEIDAE			
Lasaea australis ($2n = 36$)	5	Thiriot-Quievreux (1992)	
Lasaea colmani ($2n = 40$)	1	Ó Foighil and Thiriot-Quievreux (1999)	
Lasaea consanguinea ($2n = 100$–120)[a]	4	Thiriot-Quievreux et al. (1988b)	
Lasaea rubra ($n = 63$–340)[a]	9	Thiriot-Quievreux et al. (1989)	
Family MACTRIDAE			
Mulinia lateralis ($2n = 38$)	19	Wang and Guo (2008)	2
Family PHARIDAE			
Ensis directus ($2n = 38$)	2	Present work	1
Ensis magnus (*E. arcuatus*) ($2n = 38$)	7	Fernández-Tajes et al. (2008)	1
Ensis minor ($2n = 38$)	8	Present work	1
Ensis siliqua ($2n = 38$)	9	Fernández-Tajes et al. (2008)	1
Family SEMELIIDAE			
Scrobicularia plana ($2n = 38$)	1	Cornet and Soulard (1989)	
Family SOLECURTIDAE			
Sinonovacula constricta ($2n = 38$)	1	Wang et al. (1998)	
Family SOLENIDAE			
Solen grandis ($2n = 38$)[b]	2	Sun et al. (2003)	
Solen linearis ($2n = 38$)[b]	2	Chen et al. (2008)	
Solen marginatus ($2n = 38$)	2	Fernández-Tajes et al. (2003)	2
Family TEREDINIDIAE			
Teredo utriculus ($2n = 38$)	14	Vitturi et al. (1983)	
Family VENERIDAE			
Chamelea gallina ($2n = 38$)	4	Corni and Trentini (1986)	
Circe scripta ($2n = 38$)	3	Ebied and Aly (2004)	
Cyclina sinensis ($2n = 36$)[b]	11	Wang et al. (2001)	
Meretrix meretrix ($2n = 38$)[b]	3	Wu et al. (2002)	
Ruditapes aureus (*V. aurea*) ($2n = 38$)	1	Borsa and Thiriot-Quievreux (1990), Carrilho et al. (2011)	1
Ruditapes decussatus ($2n = 38$)	5	Ebied and Aly (2004)	
Tapes rhomboides ($2n = 38$)	4	Carrilho et al. (2011)	1
Venerupis rhomboides ($2n = 38$)	3	Insua and Thiriot-Quievreux (1992)	
Venus verrucosa ($2n = 38$)	4	Ebied and Aly (2004)	

[a] Different levels of polyploidy

[b] In Chinese; only abstract in English

T. rhomboides and *V. aurea* (Carrilho et al. 2011) with one signal. In this present work, we reported the occurrence of at least one 5S rDNA array containing a sufficient number of repeats to yield a (weak) fluorescent signal in *E. minor*. Even though in bivalve species, it is usual that some of the metaphases analysed do not yield any FISH signal (as it has previously been described by Wang et al. (2005a, b) in an oyster), the absence of a 5S rDNA signal, compared to the clear signal obtained with the major ribosomal genes probe, may well be explained by differences in the genomic organisation of these gene families. Thus, if 5S rDNA is much more dispersed in the genome of *Ensis* razor shells, compared to the major ribosomal genes (as suggested by Vierna et al. 2010), then FISH using that probe should produce a so weak signal which may probably be invisible. Differences in the genomic organisation between both species may be explained by the phylogenetic distance between American and European species, as revealed by the karyotypes (this study), shell morphology (Cosel 2009) and the ITS1-5.8S-ITS2 region (Vierna et al. 2010). The 5S rDNA multigene family is formed by a 5S rRNA coding region (corresponding to 120 nucleotides of the mature RNA) and a variable in length NTS. The 5S rDNA is characterised by a flexible organisation, as it has been found in clusters composed of similar or divergent tandemly arranged repeats (differences mainly occur in the NTS) and in clusters of 5S rDNA repeats tandemly linked to other multigene families. A dispersed organisation of 5S rDNA has also been reported, and some species were found to have more than one type of organisation within the genome (Vierna et al. 2011 and references therein). The ITS1-5.8-ITS2 and 5S rDNA regions have been studied in terms of evolutionary genetics in some *Ensis* species (Vierna et al. 2009, 2010, 2011). Vierna et al. (2010) concluded that the long-term evolution of these multigene families could be reconciled under a mixed process of concerted evolution, birth-and-death evolution and purifying selection, despite the different levels of intragenomic divergence detected (much higher within the 5S rDNA region). These authors suggested that these differences may be the consequence of a differential genomic organisation of the multigene families, that is, one or few 18S-ITS1-5.8-ITS2-28S loci containing several repeats and many 5S rDNA loci containing less repeats. Even though no conclusive data is available, our study supports their hypothesis, since the different intensities of FISH signals recorded may be explained by these differences in genomic organisation: we may have obtained weak (or none) 5S rDNA FISH signals because the repeats of this multigene family may be very dispersed within the genomes of *Ensis* species.

Finally, the hybridisation of the vertebrate telomere probe to termini of *E. minor* and *E. directus* chromosomes indicates that the vertebrate $(TTAGGG)_{22}$ sequence is present within the genomes of *E. minor* and *E. directus*, as was previously detected for the Heterodonta species *D. trunculus* (Plohl et al. 2002), *M. mercenaria* and *M. lateralis* (Wang and Guo 2001), *D. exoleta* (Hurtado and Pasantes 2005) and *T. rhomboides* and *V. aurea* (Carrilho et al. 2011).

In conclusion, this study provides new information on bivalve karyotypes, reveals important differences between American and European *Ensis* at the chromosome level, confirms *E. minor* and *E. siliqua* as separate species and supports a more dispersed organisation of the 5S rDNA compared to the major ribosomal genes.

Acknowledgments We are very grateful to Rudo von Cosel for his support on *Ensis* taxonomy. VR is supported by a "FPU" fellowship from *Ministerio de Educación y Ciencia* (Spain), and JV by a "María Barbeito" fellowship from *Xunta de Galicia* (Spain). This work was funded by the Spanish *Ministerio de Educación y Ciencia* (CTM2007-28919-E/MAR). Finally, we would like to thank two anonymous reviewers that greatly improved the quality of this article with their comments.

References

Borsa P, Thiriot-Quievreux C (1990) Karyological and allozymic characterization of *Ruditapes philippinarum*, *R. aureus* and *R. decussatus* (Bivalvia, Veneridae). Aquaculture 90:209–227

Carrilho J, Pérez-García C, Leitao A, Malheiro I, Pasantes JJ (2011) Cytogenetic characterization and mapping of rDNAs, core histone genes and telomeric sequences in *Venerupis aurea* and *Tapes rhomboides* (Bivalvia: Veneridae). Genetica 139:823–831

Chen X, Gao C, Wang J, Su Y (2008) Study on the karyotypes of *Solen linearis*. J Xiamen Univ 47:733–735

Cornet M, Soulard C (1989) Number and morphology of the metaphase mitotic chromosomes in *Scrobicularia plana* (Da Costa, 1778) (Mollusca, Bivalvia, Tellinacea). Caryologia 42:11–18

Corni MG, Trentini M (1986) A chromosomic study of *Chamelea gallina* (L.) (Bivalvia, Veneridae). Boll Zool 53:23–24

Cosel R (2009) The razor shells of the eastern Atlantic, part 2. Pharidae II: the genus Ensis Schumacher, 1818 (Bivalvia, Solenoidea). Basteria 73:1–48

Cosel R, Dörjes J, Mühlenhardt-Siegel U (1982) Die amerikanische schwertmuschel *Ensis directus* (Conrad) in der Deutschen Bucht. I. Zoogeographie und taxonomie mi vergleich mit den einheimischen schwertmuschel-Arten. Senckenberg. Marit 14:147–173

Cross I, Díaz E, Sánchez I, Rebordinos L (2005) Molecular and cytogenetic characterization of Crassostrea angulata chromosomes. Aquaculture 247:135–144

Ebied AM, Aly FM (2004) Cytogenetic studies on metaphase chromosomes of six bivalve species of families Mytilidae and Veneridae (Nucinelloidea, Mollusca). Cytologia 69:261–273

Fernández-Tajes J, González-Tizón A, Martínez-Lage A, Méndez J (2003) Cytogenetics in the razor clam *Solen marginatus* (Mollusca: Bivalvia: Solenidae). Cytogenet. Genome Res 101: 43–46

Fernández-Tajes J, Martínez-Lage A, Freire R, Guerra A, Méndez J, González-Tizón A (2008) Genome sizes and karyotypes in the razor clams *Ensis arcuatus* (Jeffreys, 1865) and *E. siliqua* (Linnaeus, 1758). Cah Biol Mar 49:79–85

González-Tizón A, Martíne-Lage A, Rego I, Ausió J, Méndez J (2000) DNA content, karyotype and chromosomal location of

18S-5.8S-28S ribosomal loci in some species of bivalve mollusc from the Pacific Canadian coast. Genome 43:409–411

Huang X, Hu X, Hu J, Zhang L, Wang S, Lu W, Bao Z (2007a) Mapping of ribosomal DNA and $(TTAGGG)_n$ telomeric sequence by FISH in the bivalve *Patinopecten yessoensis* (Jay, 1957). J Moll Stud 73:393–398

Huang X, Hu J, Hu X, Zhang G, Zhang L, Wang S, Lu W, Bao Z (2007b) Cytogenetic characterization of the bay scallop, Argopecten irradians irradians, by multiple staining techniques and fluorescence in situ hybridization. Genes Genet Sys 82:257–263

Hurtado NS, Pasantes JJ (2005) Surface spreading of synaptonemal complexes in the clam *Dosinia exoleta* (Mollusca, Bivalvia). Chromosome Res 13:575–580

Insua A, Thiriot-Quievreux C (1992) Karyotypes of *Cerastoderma edule*, *Venerupis pullastra* and *Venerupis rhomboides* (Bivalvia, Veneroida). Aquat Living Resour 5:1–8

Insua A, Freire R, Méndez J (1999) The 5S rDNA of the bivalve *Cerastoderma edule*: nucleotide sequence of the repeat unit and chromosomal location relative to 18S-28S rDNA. Genet Sel Evol 31:509–518

Iqbal ANMZ, Khan MS, Goswami U (2008) Cytogenetic studies in green mussel, *Perna viridis* (Mytiloida: Pteriomorphia), from west coast of India. Mar Biol 153:987–993

Leitao A, Chaves R, Matias D, Joaquim S, Ruano F, Guedes-Pinto H (2006) Restriction enzyme digestion chromosome banding on two commercially important veneroid bivalve species: *Ruditapes decussatus* and *Cerastoderma edule*. J Shell Res 25:857–864

Levan A, Fredga K, Sandberg AA (1964) Nomenclature for centromeric position on chromosomes. Hereditas 52:201–220

López-Piñón MJ, Insua A, Méndez J (2005) Chromosome analysis and mapping of ribosomal genes by one-and two-color fluorescent in situ hybridization in Hinnites distortus (Bivalvia: Pectinidae). J Hered 96:52–58

Martínez A, Mariñas L, González-Tizón A, Méndez J (2002) Cytogenetic characterization of *Donax trunculus* (Bivalvia: Donacidae) by means of karyotyping, fluorochrome banding and fluorescent in situ hybridization. J Moll Stud 68:393–396

Martínez-Lage A, Rodríguez-Fariña F, González-Tizón A, Prats L, Cornudella L, Méndez J (2002) Comparative analysis of different satellite DNAs in four mussel Mytilus species. Genome 45:922–929

Odierna G, Aprea G, Barucca M, Canapa A, Capriglione T, Olmo E (2006) Karyology of the Antarctic scallop *Adamussium colbecki*, with some comments on the karyological evolution of pectinids. Genetica 127:341–349

Ó Foighil D, Thiriot-Quievreux C (1999) Sympatric Australian Lasaea species (Mollusca: Bivalvia) differ in their ploidy levels, reproductive modes and developmental modes. Zool J Linn Soc 127:477–494

Pauls E, Afonso PR (2000) The karyotypes of nodipecten nodosus (Bivalvia: Pectinidae). Hydrobiologia 420:99–102

Pérez-García C, Guerra-Varela J, Morán P, Pasantes JJ (2010) Chromosomal mapping of rRNA genes, core histone genes and telomeric sequences in *Brachidontes puniceus* and *Brachidontes rodriguezi* (Bivalvia, Mytilidae). BMC Genet 11:109–117

Pérez-García C, Morán P, Pasantes JJ (2011) Cytogenetic characterization of the invasive mussel species Xenostrobus securis Lmk. (Bivalvia: Mytilidae). Genome 54:771–778

Petkeviciute R, Stunzenas V, Staneviciute G (2006) Polymorphism of the *Sphaerium corneum* (Bivalvia, Veneroida, Sphaeriidae) revealed by cytogenetic and sequence comparison. Biol J Linn Soc 89:53–64

Plohl M, Prats E, Martínez-Lage A, González-Tizón A, Méndez J, Cornudella L (2002) Telomeric localization of the vertebrate-type hexamer repeat (TTAGGG)*n*, in the wedgeshell clam *Donax trunculus* and other marine invertebrate genomes. J Biol Chem 277:19839–19846

Roiha H, Miller JR, Woods LC, Glower DM (1981) Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in *D. melanogaster*. Nature 290:749–753

Sun Z, Shao Y, Guo S, Qin Y (2003) Karyotypes of three species of marine Veneroida molluscs. Acta Oceanol Sin 22:671–678

Thiriot-Quievreux C (1992) Karyotype of *Lasaea australis*, a brooding bivalve species. Aust J Mar Freshw Res 43:403–408

Thiriot-Quievreux C, Soyer J, Bouvy M, Allen JA (1988a) Chromosomes of some subantarctic brooding bivalve species. Veliger 30:248–256

Thiriot-Quievreux C, Soyer J, Bovee F, Albert P (1988b) Unusual chromosome complement in the brooding bivalve *Lasaea consanguinea*. Genetica 75:143–151

Thiriot-Quievreux C, Insua A, Albert P (1989) Polyploidie chez un bivalve incubant, *Lasaea rubra* (Montagu). C R Acad Sci Paris 308:115–120

Thiriot-Quievreux C, Albert P, Soyer J (1990) Karyotypes of five subantarctic bivalve species. J Molluscan Stud 57:59–70

Varela MA, González-Tizón A, Francisco-Candeira M, Martínez-Lage A (2007) Isolation and characterization of polymorphic microsatellite loci in the razor clam *Ensis siliqua*. Mol Ecol Notes 7:221–222

Varela MA, Martínez-Lage A, González-Tizón A (2009) A Temporal genetic variation of microsatellite markers in the razor clam *Ensis arcuatus* (Bivalvia: Pharidae). J Mar Biol Assoc UK 89:1703–1707

Vierna J, González-Tizón A, Martínez-Lage A (2009) Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). Biochem Genet 47:635–644

Vierna J, Martínez-Lage A, González-Tizón A (2010) Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. Genome 53:23–34

Vierna J, Jensen KT, Martínez-Lage A, González-Tizón A (2011) The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae). Heredity 107:127–142

Vitturi R, Maiorca A, Catalano E (1983) The karyology *of Teredo utriculus* (Gmelin) (Mollusca, Pelecypoda). Biol Bull 165:450–457

Vitturi R, Gianguzza P, Colomba MS, Riggio S (2000) Cytogenetic characterization of *Brachidontes pharaonis* (Fisher P., 1870): karyotype, banding and fluorescence in situ hybridization (FISH) (Mollusca: Bivalvia: Mytilidae). Ophelia 52:213–220

Wang Y, Guo X (2001) Chromosomal mapping of the vertebrate telomeric sequence (TTAGGG)n in four bivalve molluscs by fluorescent in situ hybridization. J Shell Res 20:1187–1190

Wang Y, Guo X (2004) Chromosomal rearrangement in Pectinidae revealed by rRNA loci and implications for bivalve evolution. Biol Bull 207:247–256

Wang Y, Guo X (2007) Chromosomal mapping of major ribosomal rRNA genes in the hard clam (*Mercenaria mercenaria*) using fluorescence in situ hybridization. Mar Biol 150:1183–1189

Wang Y, Guo X (2008) Chromosomal mapping of the major ribosomal RNA genes in the dwarf surfclam (*Mulinia lateralis* Say). J Shell Res 27:307–311

Wang J, Zhao X, Zhou L, Xiang J (1998) Chromosome study of *Sinonovacula constricta* (Bivalvia). Oceanol Limnol Sin 29:191–196

Wang L, Xiang J, Zhou L (2001) Chromosome study of *Cyclina sinensis* Gmelin. J Northwest Sci Tech Univ Agric For 29:94–96

Wang Y, Xu Z, Guo X (2005a) Chromosomal mapping of 5S ribosomal RNA genes in the Eastern oyster, Crassostrea virginica Gmelin by fluorescence in situ hybridization. J Shell Res 24:959–964

Wang Y, Xu Z, Pierce JC, Guo X (2005b) Characterization of eastern oyster (*Crassostrea virginica* Gmelin) chromosomes by

fluorescence in situ hybridization with bacteriophage P1 clones. Mar Biotech 7:207–214
White MJD (1978) Modes of speciation. WH Freeman, San Francisco
Wu P, Dong J, Ni J, Chong J (2002) The study on chromosomes of *Meretrix meretrix*. J Shanghai Fish Univ. doi:1004-7271.0.2002-02-002
Xu Z, Guo X, Gaffney PM, Pierce JC (2001) Chromosomal location of the major ribosomal RNA genes in *Crassostrea virginica* and *C. gigas*. Veliger 44:79–83
Zhang L, Bao Z, Wang J, Wang S, Huang X, Hu X, Hu J (2007) Cytogenetic analysis in two scallops (Bivalvia: Pectinidae) by PRINS and PI banding. Acta Oceanol Sin 26:153–157
Zhang L, Bao Z, Wang S, Hu X, Hu J (2008) FISH mapping and identification of zhikong scallop (*Chlamys farreri*) chromosomes. Mar Biotech 10:151–157

4.3 Population genetic analysis of *Ensis directus* unveils high genetic variation in the introduced range and reveals a new species from the NW Atlantic

Joaquín Vierna, K. Thomas Jensen, Ana M. González-Tizón, Andrés Martínez-Lage (2012) Population genetic analysis of *Ensis directus* unveils high genetic variation in the introduced range and reveals a new species from the NW Atlantic. Marine Biology 159:2209-2227.

Bibliometrics 2012 JCR Science Edition

Impact factor: 2.468

Marine & Freshwater Biology: Q1

Mar Biol (2012) 159:2209–2227
DOI 10.1007/s00227-012-2006-6

ORIGINAL PAPER

Population genetic analysis of *Ensis directus* unveils high genetic variation in the introduced range and reveals a new species from the NW Atlantic

Joaquín Vierna · K. Thomas Jensen · Ana M. González-Tizón · Andrés Martínez-Lage

Received: 9 March 2012 / Accepted: 28 June 2012 / Published online: 31 July 2012

Abstract We report current genetic variation of populations of the razor shell *Ensis directus* (Conrad 1843) (Mollusca: Bivalvia: Pharidae) in native (North American) and introduced (European) ranges using nuclear and mitochondrial sequence-based markers. We expected less variation within the introduced range, especially considering the frequent mass mortality events observed in Europe since the species was recorded for the first time in 1978. However, we found higher variation in Europe. The possible significance of temporal fluctuations of genetic variation, limited effect of random genetic drift, and multiple introductions are discussed. Interestingly, the multiple-introduction hypothesis contrasts with the gradual colonisation of European coastal waters but is supported by trained clustering analysis and by the intensity of transatlantic shipping. Genetic and morphometric evidence strongly supports that examined individuals from a supposed *E. directus* population from Newfoundland (Canada) belong to a separate species. This new *Ensis* is formally described here and named *E. terranovensis* n.sp.

Communicated by M. I. Taylor.

Electronic supplementary material The online version of this article (doi:10.1007/s00227-012-2006-6) contains supplementary material, which is available to authorized users.

J. Vierna (✉) · A. M. González-Tizón · A. Martínez-Lage
Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, A Fraga 10, 15008 A Coruña, Spain
e-mail: jvierna@udc.es

A. M. González-Tizón
e-mail: hakuna@udc.es

A. Martínez-Lage
e-mail: andres@udc.es

J. Vierna
AllGenetics, Edificio de Servicios Centrales de Investigación, Campus de Elviña s/n, 15008 A Coruña, Spain
e-mail: joaquin@allgenetics.eu

K. T. Jensen
Marine Ecology, Department of Bioscience, Aarhus Universitet, Ole Worms Allé 1, Building 1135, 8000 Aarhus C, Denmark
e-mail: kthomas@biology.au.dk

Abbreviations

MNCN	Museo Nacional de Ciencias Naturales, Madrid, Spain
MNHN	Muséum national d'Histoire naturelle, Paris, France
ZMUC	Zoologisk Museum—Københavns Universitet, Copenhagen, Denmark
CMNML	Canadian Museum of Nature, Ottawa-Gatineau, Canada

Introduction

The razor shell *Ensis directus* (Conrad 1843) syn. *E. americanus* (Binney, in Gould and Binney 1870) (Mollusca: Bivalvia: Pharidae) is native to Atlantic North America, and it was introduced into European coastal waters in the late 1970s probably as larvae in ballast water from a ship crossing the Atlantic (von Cosel et al. 1982). The first European specimens were observed in the German Bight near the mouth of the river Elbe estuary in 1979 (von Cosel et al. 1982; Cosel 2009). Since, it has colonised the continental coastline of North Europa from Normandy to southern Norway. From 1989 onwards, it has also spread along the North Sea coast of England until the Humber

Estuary (Cosel 2009), and there have been several observations along the Channel coast. It was detected in 2002 in South Wales, near Milford Haven (Paul Dansey personal communication), and very recently, in Liverpool Bay (Dansey 2011), and in the Cantabrian Sea (Arias and Anadón 2012). Furthermore, it has also been observed in high numbers in the North Sea until a depth of about 30 m (Mühlenhardt-Siegel et al. 1983; Cosel 2009).

Ensis directus has become an integral part of the recipient ecosystem as a consumer of phytoplankton and as a prey to various fish and water birds (Tulp et al. 2010). However, no studies have been published so far that support a suppression of native species by *E. directus* in European waters. Quite the opposite, both Armonies and Reise (1999) and Dannheim and Rumohr (2011) concluded that the newcomer might have favoured the appearance of some other species in the areas in which it occurs. Nowadays, *E. directus* is a commercial species in Europe (see Marine Stewardship Council 2012). Similarly, in the north eastern US and eastern Canada, there is an increasing interest in *E. directus* fisheries and valuable work was carried out to investigate the aquaculture potential of the species in Maine (Maine Sea Grant 2012) and in Nova Scotia (Kenchington et al. 1998).

The dispersal of *E. directus* is facilitated by a pelagic larval life (2–4 weeks) during which it may reach a distance of 125 km downstream from its source population (Armonies 2001) (estimate based on its dispersal from 1979). Human-facilitated spread within its new ecosystem could be a further dispersal mechanism that may have contributed to its present European distribution. According to Armonies (2001), the temporal course of dispersal along the coastline fits the hypothesis of a single introduction into the North Sea but the possibility of multiple imports of the species cannot be excluded, and this may be tested genetically.

One important factor impacting the genetic variation of the European populations could be the often observed mass mortalities in local *E. directus* populations that seem to be a characteristic feature of the species. For instance, thousands of dead or dying razor shells have been observed several times along the Danish shoreline (North Sea, Limfjorden, and Kattegat areas) (Freudendahl et al. 2010). Such events may reduce levels of genetic variation either because of selective mortality or through random eradication of genotypes.

In population genetics, the analysis of several loci has become a must, as studies based on nucleotide variation at a single locus provide insufficient information of genetic patterns. In this sense, the selection of loci with different evolutionary histories (e.g. cytoplasmic vs nuclear genes) is important to distinguish among factors affecting genetic variation (e.g. low variation due to a selective sweep or a population bottleneck). But, the selection of markers is restricted to availability of suitable primers in non-model organisms.

In this work, we have studied nucleotide variation at four sequence-based molecular markers, including mitochondrial, nuclear multi-copy, and nuclear single-copy regions in order to: (1) compare current genetic variation in native and introduced ranges to assess to what extent potential bottlenecks and mass mortality events in Europe have impacted diversity, (2) obtain information about the possible origin of European individuals, and (3) study population structure in the native range of the species. Besides, we serendipitously detected a very divergent population from Conception Bay, Newfoundland (Canada) that is proposed to be a new *Ensis* species based on genetic and morphometric evidence. This new *Ensis* species from the NW Atlantic is formally described here.

Materials and methods

Specimens and lab procedures

We studied a set of 148 razor shells from native and introduced sites (see Table 1; Fig. 1 for details). Razor shells were preserved in 100 % ethanol, and they were identified in the lab as *E. directus* according to shell morphology (Cosel 2009). Identifications of some of the specimens (Table 1) were confirmed by Rudo von Cosel (Muséum national d'Histoire naturelle, Paris, France) and included in his article on the taxonomy of Atlantic *Ensis* (Cosel 2009). One individual from the related species *E. minor* Dall, 1899 (collected off Christmas Bay, Texas, USA) was included as outgroup in some of the analyses.

DNA was extracted from muscle tissue using the NucleoSpin Tissue kit (Macherey–Nagel GmbH and Co. KG). All razor shells were sequenced for a fragment of the mitochondrial cytochrome oxidase subunit I gene (COI), and for the nuclear ribosomal multi-copy region encoding both internal transcribed spacers (ITS1 and ITS2) and the 5.8S ribosomal gene (5.8S). A subset of 70 animals (including the outgroup) were additionally sequenced for a fragment of a nuclear single-copy region, the adenine nucleotide translocase gene (ANT). Using the 'universal' primers designed by Folmer et al. (1994) and Audzijonyte and Vrijenhoek (2010), we obtained sequences from three individuals that were then used to design three pairs of internal primers in GeneFisher (Giegerich et al. 1996) (COI-*directus*-F, 5′ CAG GTT TAG TTG GAA CTA GG; COI-*directus*-R, 5′GAT CTC CRC CAC CTC T; ANT-*Ensis*-a-F, 5′ AAA CAT GGC CAA CTG CAT CCG AT; ANT-*Ensis*-a-R, 5′ CAA GGA CAT AAA GCC CTC TGC CTT; ANT-*Ensis*-b-F, 5′ TTC CCA ACC CAG GCC TTG;

Table 1 Sites sampled in this work

Sampling site	Coordinates	Collected by	Years	Depth	Identified by	Museum code
Sillerslev	56°42′20″N, 8°47′20″E	K. T. Jensen	2005	1–2 m	K. T. Jensen	
Sundsøre	56°42′25″N, 9°10′31″E	K. T. Jensen	2005	1–2 m	K. T. Jensen	
Juvre Deep	55°11′45″N, 8°25′56″E	K. T. Jensen	2005	Intertidal (<2 m)	K. T. Jensen	
Vester Vedsted	55°16′29″N, 8°37′45″E	K. T. Jensen	2008	Intertidal (<2 m)	J. Vierna	
The Wash	52°56′21″N, 0°24′53″E	D. Palmer	2007	1–2 m	R. von Cosel, J. Vierna	
Katwijk	52°11′57″N, 4°24′41″E	J. Goud	2008	Thrown on beach after storm (ca. 1–3 m)	J. Goud, J. Vierna	RMNH.MOL.102103
Dunkerque	51°02′03″N, 2°16′37″E	J. M. Dewarumez	2009	Intertidal (<2 m)	J. Vierna	
Cobscook Bay	44°54′35″N, 67°4′13″W	T. Sheehan (Gulf of Maine)	2008	1.8 m at high mean water (intertidal)	J. Vierna	MNCN 15.07/11733
Shinnecock Bay	40°52′07″N, 72°28′02″W	S. T. Tettelbach	2008	0.5–3 m	J. Vierna	MNCN 15.07/11734
Long Pond	47°30′54″N, 52°58′30″W	P. Sargent, R. O'Donnell	2007	6–11 m	R. von Cosel, J. Vierna	Several codes (see 'Taxonomy')

RMNH, NCB Naturalis, Leiden (The Netherlands). *MNCN* Museo Natural de Ciencias Naturales, Madrid (Spain)

ANT-*Ensis*-b-R, 5′ ATG ATG GTY GTG GCA CAG T). These internal primers were used in all subsequent amplifications. The primers used to amplify the ITS1-5.8S-ITS2 region were those from Heath et al. (1995). Each PCR reaction (25 μL) contained ~25 ng of genomic DNA, 0.625 U of *Taq* DNA polymerase (Roche Diagnostics), 5 nmol of each dNTP (Roche Diagnostics), 20 pmol of each primer, and the buffer recommended by the polymerase supplier. The general reaction conditions were: an initial denaturation step at 94 °C for 3 min followed by 35 cycles of denaturation at 94 °C for 20 s; annealing at the following temperatures (COI-*directus*, 48 °C; ANT-*Ensis*-a, 58 °C; ANT-*Ensis*-b, 54 °C; and ITS1-5.8S-ITS2 region, 59 °C) for 20 s; extension at 72 °C for 30–50 s; and a final extension at 72 °C for 5 min. PCR products were run on agarose gels, stained with ethidium bromide, and imaged under UV light. COI and ANT PCRs yielded one single gel band from the expected sizes, and the amplification products were purified using ExoSAP-IT (USB). COI amplification products were sequenced in both directions using PCR primers, whereas ANT amplicons were sequenced using the ANT-*Ensis*-b-F primer only. ITS1-5.8S-ITS2 PCRs yielded single-band patterns. Nevertheless, since intragenomic variants occur in *E. directus* (Vierna et al. 2010), a cloning step was necessary. Therefore, amplification products were cloned using the TOPO TA Cloning kit (Invitrogen). We selected transformant colonies, checked their insert size by PCR, spread one clone per individual on an LB plate, and let it grow overnight at 37 °C. Plasmids were purified with the QiaPrep Spin Miniprep Kit (Qiagen), and they were sequenced using the M13 Forward and Reverse primers supplied in the cloning kit.

Bioinformatic analyses

The software BioEdit 7.0.9.0 (Hall 1999) and Geneious Pro 5.4.6 (Drummond et al. 2011) were used to examine the electropherograms. Since we expected heterozygote positions in the ANT sequences, we used Geneious Pro 5.4.6 to detect them (peak similarity = 50 %). Afterwards, they were confirmed manually. The 5.8S region was not further considered here because it was almost invariable. Alignments were carried out in ClustalW 2.0 (Larkin et al. 2007) from the MEGA 5.03 package (Tamura et al. 2011). In the case of COI and ANT, they were performed considering the amino acid sequence. ITS1 and ITS2 alignments were manually corrected using the RALEE (RNA ALignment Editor in Emacs) tool (Griffiths-Jones 2005).

The gametic phase of each ANT sequence was obtained in DnaSP 5.10.01 (Librado and Rozas 2009) implementing the algorithm provided in PHASE 2.1 (Stephens et al. 2001; Stephens and Scheet 2005). PHASE simulations were run five times (with a different seed each time) using the MR0 model and assuming recombination. Output probability thresholds were 0.9, and all other parameters

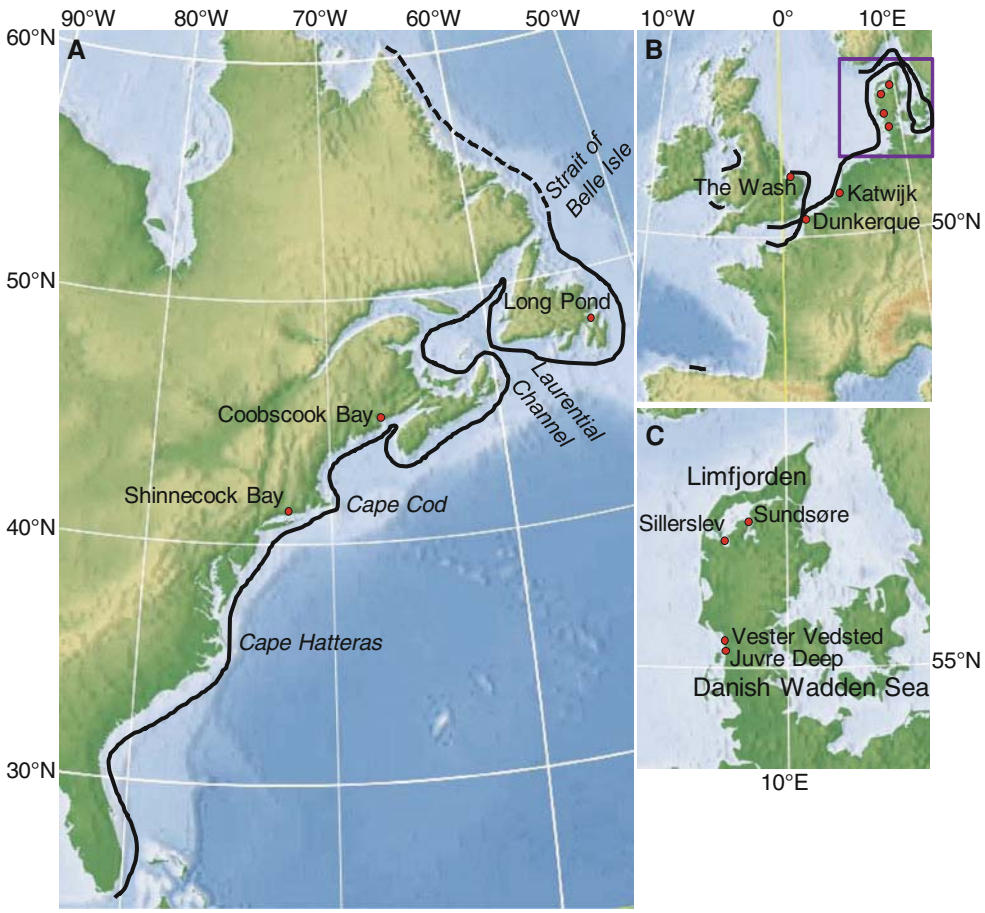


Fig. 1 Maps showing the approximate current distribution of *E. directus* according to Kenchington et al. (1998), Cosel (2009), Dansey (2011), and Arias and Anadón (2012). *Red dots* correspond to sampling sites. In the *dashed lined* area, the occurrence of the species is unclear. Potential barriers to gene flow in the Western Atlantic Boreal Region are indicated. **a** Native range. **b** Introduced range. **c** Introduced range (detail)

were set as default. No differences were appreciated among the five resulting alignments (140 sequences, each), which were inspected for discrepancies by comparing their consensus sequences (constructed in Geneious Pro 5.4.6), and their nucleotide diversity values (obtained from DnaSP 5.10.01).

Diversity analyses were done in DnaSP 5.10.01, after creating several subsets according to sampling sites or geographic areas. In the cases of ITS1 and ITS2 sequences, gap-containing positions were excluded in each subset. However, alignment gaps were considered in the haplotype/sequence-type data files that we created in DnaSP 5.10.01 (which are compatible with the Arlequin 3.5.1.2 software, Excoffier and Lischer 2010). Therefore, for the over-mentioned datasets, two values for the number of sequence-types were provided, one excluding gaps that were present in the subset, and one including all gaps.

In Arlequin 3.5.1.2, we looked for shared haplotypes/sequence-types among sampling sites. Using this software, we also performed analyses of molecular variance (AMOVAs, Excoffier et al. 1992), in which sampling sites were grouped in different ways, and running 1,000 permutations. The Tamura—Nei distance was employed to build up the matrix. Danish sites were clustered in pairs considering their geographic proximity: Sillerslev and Sundsøre (Limfjorden); Juvre Deep and Vester Vedsted (Danish Wadden Sea). The fixation index F_{ST} was also calculated for each pair of population comparisons, and its significance was tested by 10,000 permutations. Finally, exact tests of population differentiation were carried out. They test the hypothesis of random distribution of individuals between pairs of populations as described in Raymond and Rousset (1995) and Goudet et al. (1996). The number of steps in Markov chain was set to 1,000,000, and the number of dememorisation steps, to 100,000.

In the F_{ST} and test of population differentiation analyses, we considered only four groups of sequences (Europe, Shinnecock Bay, Cobscook Bay, and Long Pond) following Fitzpatrick et al. (2012). Since these methods require populations to have reached a mutation-drift equilibrium, the authors pointed out that such population genetic methods must not be applied to study population structure within a species' introduced range. They claimed that the time since the initial introduction and subsequent expansion of range of an introduced species is too recent for the effects of mutation and drift to be at equilibrium, particularly when effective population sizes are large (as in marine bivalves). Only regarding these comparisons and the

neutrality tests (see below), we assume that European sequences represent the genetic variation of their source population. This would only be the case when no genetic bottlenecks occur, the source population is only one, and the introduction event is only one as well.

In order to investigate whether there was a native population more likely to be the source of European individuals, we input our COI-ITS1-ITS2 dataset in BAPS 5.3 (Corander et al. 2006, 2008) and ran a trained clustering analysis (Cheng et al. 2011) with a maximum number of clusters ranging between $K = 3$ and $K = 9$. The ANT sequences were not used because they were not available for all individuals. The analysis was repeated five times with invariable results. Individuals from Long Pond were excluded from this analysis because they did not share haplotypes/sequence-types with European individuals (see below).

Several neutrality tests were calculated with COI and ANT datasets. The ITS1 and ITS2 sequences were excluded since the occurrence of intragenomic divergence within this region in *E. directus* (Vierna et al. 2010) could violate some test assumptions and produce biased results. Tajima's *D* (Tajima 1989) and Fu's *Fs* (Fu 1997) were calculated in Arlequin 3.5.1.2, applying a number of simulated samples of 1,000, and Fu and Li's *D* and *F* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000), in DnaSP 5.10.01, using the total number of mutations, and *E. minor* as outgroup. The McDonald-Kreitman test (McDonald and Kreitman 1991) was also performed in DnaSP 5.10.01. In this case, the Canadian sequences were used as outgroup for each European partition, and the European sequences (all together) were used as outgroup for the Canadian sequences. The Hudson–Kreitman–Aguadé test (Hudson et al. 1987) was implemented for the COI + ANT datasets, taking *E. minor* sequences for interspecific comparisons, and using the 'direct mode', in DnaSP 5.10.01. This mode allows to compare loci that differ in their effective population sizes.

Mismatch distributions of the pairwise number of differences (Rogers and Harpending 1992) were obtained, and the goodness of fit of the observed and expected curves (under the 'sudden expansion' model) was assessed by the sum of squared deviations (SSD) and the raggedness statistic (Harpending 1994). The Θ_0, Θ_1, and τ parameters were calculated in Arlequin 3.5.1.2 with 100 bootstrap replicates, and the output values were introduced in DnaSP 5.10.01, where the histograms for each partition were obtained under a 'population growth-decline' model.

Mismatch analyses were also used to roughly estimate the time elapsed since expansion of the partitions that did not show significant SSD/raggedness statistic tests (i.e. those that may have undergone an expansion event). We used the equation $\tau = 2ut$ (Rogers and Harpending 1992), where $u = 2\ \mu l$, being μ the number of mutations per nucleotide site per generation, and l, the sequence length. Generation time was assumed to be 5 years, though *E. directus* reaches sexual maturity after 1 year (see Mühlenhardt-Siegel et al. 1983). Mutation rates for COI (0.14 and 0.52 % divergence per nucleotide site per million years) were obtained from Luttikhuizen et al. (2003) who studied another heterodont species. We are not aware of any reported estimate of mutation rates for the ANT region.

COI and ANT networks were calculated under the maximum parsimony criterion in TCS 1.21 (Clement et al. 2000) using the haplotype/sequence-type dataset, and applying a connexion limit high enough to permit the outgroup connexion. Due to their higher complexity, ITS1, ITS2, and ITS1 + ITS2 networks were built up from the sequence-type datasets using the neighbournet algorithm (Bryant and Moulton 2004) and uncorrected p-distances in SplitsTree4 (Huson and Bryant 2006).

Phylogenies inferred under maximum likelihood (ML), bayesian (BA), and maximum parsimony (MP) criteria were obtained for the COI + ANT + ITS1 + ITS2 concatenated dataset (which comprised sequences from 69 individuals and 1,710 nucleotides). RAxML-7.2.8 (Stamatakis 2006, 2008) was run from the CIPRES Science Gateway (Miller et al. 2010). This software is capable to assign and estimate separate model parameters for individual genes of multi-gene alignments (Stamatakis 2006) and implements the general time reversible (GTR) substitution model for all partitions. Therefore, we partitioned our data by genes. A GTRCATI model of nucleotide substitution was implemented, and 1000 automatic bootstraps were performed, followed by a search for the best-scoring ML tree. Gaps were considered as missing data. An additional ML analysis was performed in PhyML (Guindon et al. 2010) that was run through the ATGC Bioinformatics Platform (http://www.atgc-montpellier.fr/), implementing a GTR model of nucleotide substitution, 1,000 non-parametric bootstraps, and considering gaps as missing data. All other parameters were set to their default values, and in this case, data were not partitioned by genes. BA was carried out using the software MrBayes v. 3.0B4 (Huelsenbeck and Ronquist 2001) through the CIPRES Science Gateway. Models of evolution for each partition were obtained from MrModelTest v2.3 (Johan Nylander, http://www.abc.se/~nylander/). This software selected the HKY + I + G model for the COI partition, the K80 model for ANT, and the GTR + I + G model for either ITS. Gaps were treated as missing data. The analysis was performed with 15,000,000 generations initiated with a random starting tree, sampling every 1,000 generations and allowing the program to estimate the likelihood parameters required. Stationarity was assessed using the web-based software AWTY (Nylander et al. 2008). Results collected prior to

stationarity were discarded as burn-in. A MP bootstrap consensus tree was retrieved from Paup4.0b10 (Swofford 2002) using the heuristic search method. Parameters were set as follows: gaps were treated as a 'fifth state' (this applies to the ITS1 and ITS2 sequences), multistate taxa were interpreted as uncertainty (this applies to the ANT sequences), starting trees were obtained via stepwise addition, the number of trees held at each step during stepwise addition was one, and the branch-swapping algorithm selected was TBR. The robustness of the obtained topology was assessed after running 1,000 non-parametric bootstraps. All phylogenetic trees were edited in FigTree 1.2.2. (Andrew Rambaut, http://tree.bio.ed.ac.uk/software/figtree/).

Finally, to compare our results to other reported cases, between-groups mean K2P distances were calculated using the MEGA 5.03 package. In the case of ITS1 and ITS2, gaps were considered in pairwise comparisons. Standard errors were obtained after running 1,000 permutations.

Morphometric analyses

Razor shells from the three native sites were additionally studied in terms of shell morphometrics. Two individuals from Cobscook Bay and two other from Long Pond that were studied genetically could not be studied morphologically because their valves were missing.

Using a vernier calliper, we measured the length of each right valve (considering the longest axis), and the width at the posterior adductor scar. The length of the posterior adductor scar (named here 'distance a'), and the distance between its posterior end and the beginning of the pallial sinus (along the dorsal line) ('distance b') were also recorded, since they seemed to differ among sites. Furthermore, the shape of muscle scars is a main taxonomic character in *Ensis* (see Cosel 2009). Distances measured are indicated in Fig. 2. We also weighed each pair of valves (whenever possible, since some of the shells were incomplete) and obtained the mean value for each valve.

To analyse morphometric data, we used the statistical programme SPSS vs 19. Prior to running ANCOVA, regression residuals were tested for homoscedasticity.

Results

Genetic variation

COI, ITS1, and ITS2 sequences were obtained for 77 European and 71 North American individuals. In addition, we obtained ANT sequences for a subsample of 45 European and 24 North American individuals. These sequences were converted into 90 European and 48 North American phased sequences.

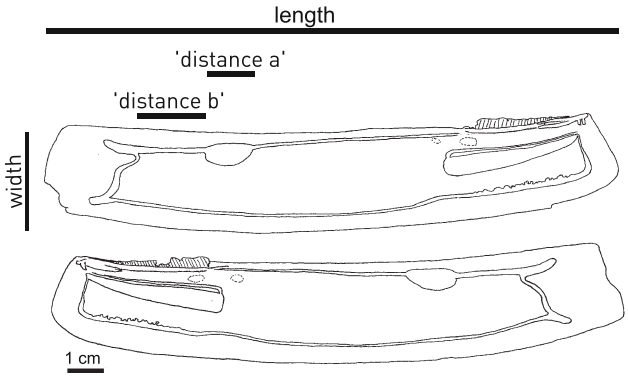


Fig. 2 Internal view of the valves of the holotype of *Ensis terranovensis* n.sp., showing the muscle scars and the distances considered in the morphometric analyses. '*Distance a*', length of the posterior adductor scar; '*distance b*', distance along the dorsal line between the posterior end of the posterior adductor scar and the beginning of the pallial sinus

After having deleted the primer-annealing regions and some low-quality terminal regions of the electropherograms, we obtained sets of sequences with the following lengths: COI, 454 bp; ITS1, 484–514 bp; ITS2, 295–301 bp; and ANT, 406 bp. All alignments were straightforward, despite some minor corrections performed in the ITS1 and ITS2 datasets. All sequences can be accessed at the DDBJ/EMBL/GenBank databases (accession numbers HE661632–HE662148).

Polymorphism values for each marker are indicated in Table 2 by site and geographic area. COI and ITS1 were more polymorphic than ITS2 and ANT. Considering all sequences (except the outgroup), COI displayed the highest proportion of segregating sites ($s = 0.165$), and the highest nucleotide diversity ($\Pi = 0.025 \pm 0.002$) and average number of nucleotide differences ($k = 11.303$). The values for the ITS1 region were higher in terms of the number of sequence-types obtained ($h = 73$; $h = 111$, depending on gaps, as explained above) and in terms of haplotype diversity ($H = 0.963 \pm 0.008$).

Strikingly, European sequences were, in general, more polymorphic than those from native sites (Table 2). Only the H value for Cobscook Bay ITS2 and for Shinnecock Bay ITS1 and the Π value for Long Pond COI sequences were higher in native populations. In the particular case of ITS1 and ITS2, European, Shinnecock Bay, and Cobscook Bay sequences were more polymorphic than those from Long Pond, but this was probably due to the occurrence of paralog ITS sequences in the genomes of these razor shells, which were not present in Long Pond individuals (see 'Evolutionary relationships among variants').

Table 2 Polymorphism in sampling sites or geographic areas

	Limfjorden		Danish Wadden Sea		The Wash	Katwijk	Dunkerque	European seq.	Shinnecock Bay	Cobscook Bay	Long Pond	*N* (total)
	Sillerslev	Sundsøre	Juvre Deep	Vester Vedsted								
COI												
N	8	8	8	8	15	15	15	77	21	22	28	148
l	454	454	454	454	454	454	454	454	454	454	454	454
S	20	16	15	13	18	15	18	43	21	12	20	75
s	0.044	0.035	0.033	0.029	0.040	0.033	0.040	0.095	0.046	0.026	0.044	0.165
h	7	7	7	6	11	9	10	36	12	7	15	62
H	0.964 ± 0.077	0.964 ± 0.077	0.964 ± 0.077	0.893 ± 0.111	0.952 ± 0.04	0.914 ± 0.052	0.914 ± 0.056	0.948 ± 0.013	0.905 ± 0.048	0.792 ± 0.066	0.884 ± 0.045	0.949 ± 0.011
Π	0.012 ± 0.001	0.012 ± 0.002	0.011 ± 0.002	0.008 ± 0.002	0.010 ± 0.001	0.009 ± 0.001	0.010 ± 0.001	0.010 ± 0.000	0.008 ± 0.001	0.006 ± 0.001	0.013 ± 0.009	0.025 ± 0.002
k	5.643	5.536	5.036	3.536	4.343	4.19	4.438	4.678	3.571	2.736	5.738	11.303
ITS1												
N	8	8	8	8	15	15	15	77	21	22	28	148
l	495	498	492	494	494	495	495	487	495	493	484	530
S	18	26	20	22	21	14	18	45	20	21	7	62
s	0.036	0.052	0.041	0.045	0.043	0.028	0.036	0.092	0.040	0.043	0.014	0.117
h	8 (8)	8 (8)	8 (8)	8 (8)	12 (14)	13 (14)	14 (15)	48 (64)	15 (20)	19 (19)	7 (9)	73 (111)
H	1.000 ± 0.063	1.000 ± 0.063	1.000 ± 0.063	1.000 ± 0.063	0.943 ± 0.054	0.981 ± 0.031	0.990 ± 0.028	0.964 ± 0.012	0.952 ± 0.032	0.983 ± 0.021	0.791 ± 0.041	0.963 ± 0.008
Π	0.014 ± 0.002	0.017 ± 0.005	0.015 ± 0.002	0.016 ± 0.002	0.010 ± 0.002	0.011 ± 0.001	0.011 ± 0.001	0.011 ± 0.001	0.009 ± 0.001	0.011 ± 0.001	0.004 ± 0.001	0.014 ± 0.001
k	7.179	8.500	7.357	7.857	5.143	5.314	5.676	5.292	4.533	5.632	2.013	6.794
ITS2												
N	8	8	8	8	15	15	15	77	21	22	28	148
l	295	295	295	295	295	293	294	292	294	295	295	320
S	9	8	10	9	10	10	12	18	12	12	4	25
s	0.031	0.027	0.034	0.031	0.034	0.034	0.041	0.062	0.041	0.041	0.014	0.078
h	6 (6)	7 (7)	7 (7)	6 (6)	12 (12)	8 (10)	10 (11)	22 (29)	12 (16)	13 (14)	5 (5)	31 (50)
H	0.929 ± 0.084	0.964 ± 0.077	0.970 ± 0.077	0.929 ± 0.084	0.962 ± 0.040	0.895 ± 0.003	0.933 ± 0.002	0.929 ± 0.012	0.933 ± 0.031	0.909 ± 0.043	0.725 ± 0.042	0.918 ± 0.009
Π	0.014 ± 0.002	0.010 ± 0.003	0.016 ± 0.002	0.016 ± 0.002	0.014 ± 0.001	0.011 ± 0.002	0.015 ± 0.002	0.014 ± 0.001	0.013 ± 0.001	0.013 ± 0.001	0.004 ± 0.000	0.014 ± 0.001
k	4.143	2.893	4.857	4.607	4.21	3.238	4.343	4.018	3.943	3.991	1.336	4.081
ANT												
N		4	8	10	24	24	20	90	6	12	30	138
l		406	406	406	406	406	406	406	406	406	406	406
S		3	2	3	9	7	7	16	2	5	5	22
s		0.007	0.005	0.007	0.022	0.017	0.017	0.039	0.005	0.012	0.012	0.054
h		3	3	3	11	9	8	17	4	5	6	24
H		0.833 ± 0.222	0.607 ± 0.164	0.622 ± 0.138	0.848 ± 0.063	0.855 ± 0.051	0.805 ± 0.07	0.821 ± 0.029	0.8 ± 0.172	0.742 ± 0.116	0.634 ± 0.08	0.860 ± 0.020
Π		0.004 ± 0.001	0.002 ± 0.001	0.003 ± 0.001	0.004 ± 0.001	0.004 ± 0.000	0.004 ± 0.001	0.004 ± 0.000	0.003 ± 0.001	0.003 ± 0.001	0.003 ± 0.000	0.004 ± 0.001
k		1.500	0.679	1.067	1.558	1.518	1.558	1.432	1.067	1.348	1.067	1.630

European sequences are those from Sillerslev, Sundsøre, Juvre Deep, Vester Vedsted, The Wash, Katwijk, and Dunkerque. *N* number of sequences considered in each analysis, *l* length of the genomic region studied, *S* number of segregating sites, *s* number of segregating sites per nucleotide site, *h* number of haplotypes or sequence-types, *H* haplotype diversity, Π nucleotide diversity, *k* average number of nucleotide differences, *h* values in brackets are those in which gaps were considered (see 'Materials and methods'). *H* and Π values are expressed with their SD

Distribution of variants

A graphic representation of the distribution of all variants is available (Online Resource 1). We sampled 62 COI haplotypes (out of 148 sequences), and 36 of them were found in the introduced range. The number of private European haplotypes was 30, and the remaining six were shared with Shinnecock Bay (four haplotypes), Cobscook Bay (four haplotypes) or both sites (two haplotypes). Shinnecock Bay displayed 12 COI haplotypes, eight of which were private; and Cobscook Bay, seven haplotypes (three of them were private). The number of shared variants between Shinnecock Bay and Cobscook Bay was only two. In Long Pond, we sampled 15 haplotypes, all of them, private.

In the case of ITS1, 111 sequence-types were sampled, 63 of them were private to Europe, and one was shared between Europe and Cobscook Bay. Shinnecock Bay displayed 20 (all private) variants; and Cobscook Bay, 18 privates and the one shared with Europe. There were no shared variants between Shinnecock Bay and Cobscook Bay. Long Pond displayed nine variants, all private.

The number of ITS2 sequence-types obtained was 50, 29 of them occurring in Europe. Among those, 20 variants were private to this region, eight were shared with Shinnecock Bay, five with Cobscook Bay, and four with both sites. In Shinnecock Bay, we sampled 16 ITS2 sequence-types (seven were private); and in Cobscook Bay, 14 sequence-types (eight, private). The number of shared variants between Shinnecock Bay and Cobscook Bay was five. Again, all Long Pond sequence-types (five) were private to this sampling site.

All sampling sites displayed many private sequence-types when the ITS1 + ITS2 dataset was considered (see Online Resource 1).

Finally, the ANT region was more conserved, and the number of sequence-types sampled was 24 (in this case, out of 138 sequences); 13 of them were private to European waters, and four sequence-types were shared among European and North American sampling sites. In Shinnecock Bay, we sampled four sequence-types (one private, three shared with Europe and with Cobscook Bay, one shared with Long Pond). In Cobscook Bay, we found five (one private, three shared with Shinnecock Bay, four with Europe, one with Long Pond). In Long Pond, we found six (five private, one shared with Europe, Shinnecock Bay, and Cobscook Bay).

Population differentiation

AMOVA results showed that, when only introduced sampling sites were considered (Limfjorden–Danish Wadden Sea–The Wash–Katwijk–Dunkerque), the percentage of genetic variation within populations was very high (97.51 % for COI, 101.50 % for ITS1, 97.18 % for ITS2, and 99.21 % for ANT).

Similarly, in the comparison European sites (altogether)—Cobscook Bay—Shinnecock Bay, the percentage of genetic variation within populations was 95.31 % for COI, 100.51 % for ITS1, 101.71 % for ITS2, and 99.91 % for ANT, again indicating a lack of structure.

There was one grouping that maximised the percentage of variation among groups. It was the one in which Long Pond was separated from all other sampling sites (European sequences + Shinnecock Bay + Cobscook Bay − Long Pond). Specifically, the values obtained were: COI, 83.64 %; ITS1, 63.91 %; ITS2, 60.03 %; and ANT, 39.57 %.

F_{ST} values were significant in the COI comparisons between European sequences and either Shinnecock Bay ($F_{ST} = 0.03$) or Cobscook Bay ($F_{ST} = 0.03$). These values suggest low differentiation according to mitochondrial DNA. Tests of population differentiation were significant in the comparisons between European sequences and either Shinnecock Bay or Cobscook Bay according to ITS1.

In the comparisons between Shinnecock Bay and Cobscook Bay, F_{ST} values were non-significant, but the COI test of population differentiation suggested some degree of differentiation between these native sites.

Finally, both types of tests and all four molecular markers yielded significant results in all comparisons between Long Pond and all other sites or areas (Table 3).

Most likely source population of European individuals

According to BAPS 5.3 analyses (Online Resource 2), the number of groups in the optimal partition was two, that is, the software did not need additional source populations to explain the diversity of European individuals. BAPS 5.3 assigned a probability to each European individual of belonging to a particular cluster (in this case, Shinnecock Bay or Cobscook Bay). Among European individuals, 49 out of 77 (63.6 %) were more likely to belong to the Cobscook Bay cluster, and the remaining 28 (36.4 %) were more likely to belong to Shinnecock Bay. Interestingly, a high number of individuals were more likely to belong to the Cobscook Bay cluster in the Limfjorden (75 %), Danish Wadden Sea (75 %), and Katwijk (73 %) areas. On the contrary, The Wash and Dunkerque individuals were more balanced (53.3 % were more likely to belong to the Shinnecock Bay cluster, and the remaining 46.7 %, to Cobscook Bay).

Neutrality tests and past changes in population size

Results of all tests performed are recorded in Table 4. Only the Fu's *Fs* test was significant in one data partition

Table 3 Population differentiation

	European seq.	Shinnecock Bay	Cobscook Bay	Long Pond
European seq.				
COI	–	ITS1	ITS1	COI, ITS1, ITS2, ANT
ITS1	–			
ITS2	–			
ANT	–			
Shinnecock Bay				
COI	***0.03***	–	COI	COI, ITS1, ITS2, ANT
ITS1	0.01	–		
ITS2	−0.02	–		
ANT	0.02	–		
Cobscook Bay				
COI	***0.03***	0.01	–	COI, ITS1, ITS2, ANT
ITS1	0.01	0.03	–	
ITS2	−0.02	−0.02	–	
ANT	−0.01	−0.07	–	
Long Pond				
COI	***0.83***	***0.83***	***0.84***	–
ITS1	***0.64***	***0.74***	***0.67***	–
ITS2	***0.48***	***0.55***	***0.56***	–
ANT	***0.39***	***0.45***	***0.42***	–

European sequences are those from Sillerslev, Sundsøre, Juvre Deep, Vester Vedsted, The Wash, Katwijk, and Dunkerque. F_{ST} values are below diagonal (values in bold italics are significant). Above diagonal, tests of population differentiation that resulted to be significant for each marker. Significance level, $\alpha = 0.05$

Table 4 Neutrality tests

	European seq.	Shinnecock Bay	Cobscook Bay	Long Pond
COI				
N	77	21	22	28
Tajima's *D*	−1.506 ($P = 0.032$)	−1.464 ($P = 0.056$)	−0.592 ($P = 0.308$)	0.411 ($P = 0.690$)
Fu's *Fs*	***−22.784 (P = 0.000)***	−3.977 ($P = 0.037$)	−0.060 ($P = 0.528$)	−3.002 ($P = 0.118$)
Fu and Li's *D*	−1.949 (0.1 > P>0.05)	−0.470 ($P > 0.1$)	0.126 ($P > 0.1$)	−1.140 ($P > 0.1$)
Fu and Li's *F*	−2.235 (0.1 > P>0.05)	−1.007 ($P > 0.1$)	−0.325 ($P > 0.1$)	−1.073 ($P > 0.1$)
Fay and Wu's *H*	−18.064	−11.048	−8.537	−2.365
MK (Fisher's exact test)	$P = 0.354$	nct	nct	$P = 0.354$
ANT				
N	90	6	12	30
Tajima's *D*	−1.549 ($P = 0.042$)	1.032 ($P = 0.853$)	−0.684 ($P = 0.287$)	−0.422 ($P = 0.413$)
Fu's *Fs*	***−10.415 (P = 0.000)***	−1.685 ($P = 0.021$)	−1.159 ($P = 0.143$)	−1.454 ($P = 0.201$)
Fu and Li's *D*	−1.135 ($P > 0.1$)	0.883 ($P > 0.1$)	−1.553 ($P > 0.1$)	0.244 ($P > 0.1$)
Fu and Li's *F*	−1.712 ($P > 0.1$)	1.005 ($P > 0.1$)	−1.783 ($P > 0.1$)	0.051 ($P > 0.1$)
Fay and Wu's *H*	−1.113	0.267	0.697	−1.177
MK (Fisher's exact test)	nct	nct	nct	nct
COI + ANT				
HKA test (direct mode)	$P = 0.895$	$P = 0.418$	$P = 0.778$	$P = 0.697$

European sequences are those from Sillerslev, Sundsøre, Juvre Deep, Vester Vedsted, The Wash, Katwijk, and Dunkerque. Significance of some of the tests was evaluated by means of the *P* value (*P*). Significant values after the Bonferroni correction ($\alpha = 0.0125$) are bold italic. *MK* McDonald-Kreitman test, *nct* the contingency table could not be computed by the software because there were no fixed differences between data sets

Table 5 Mismatch distribution parameters and estimates of time elapsed since expansion in the European partition

	COI	ANT
N	77	90
k	4.678	1.432
Observed variance	4.340	0.964
Θ_0	0.000	0.002
Θ_1	83.438	99999
τ	5.318	1.574
SSD	0.008 ($P = 0.06$)	0.012 ($P = 0.07$)
Harpending's RI	0.021 ($P = 0.18$)	***0.100 (P = 0.01)***
TESE (years)		
μ (%)	**0.14**	20917244
μ (%)	**0.52**	5631566

European sequences are those from Sillerslev, Sundsøre, Juvre Deep, Vester Vedsted, The Wash, Katwijk, and Dunkerque. Significance of the SSD and Harpending's RI tests were evaluated by means of the *P* value (*P*). Values in bold italics and bold are significant ($\alpha = 0.05$). *k* average number of nucleotide differences, *TESE* time elapsed since expansion, *SSD* sum of squared deviations, *Harpending's RI* Harpending's raggedness index

(European sequences) after applying the Bonferroni correction. It was significantly negative in the COI ($Fs = -22.784$) and ANT ($Fs = -10.415$) datasets, suggesting recent population expansion or genetic hitchhiking. Since both markers showed congruent results, and considering that it is not feasible that selection had been acting both over the nuclear and the mitochondrial regions under study (see Beaumont 2007), the past population expansion hypothesis is more likely.

The HKA test did not reveal significant deviations from neutrality of any of these two markers in any partition, meaning that the genomic regions chosen seem to have evolved neutrally.

Mismatch distributions of the pairwise number of differences were obtained for the European partition (the one in which we detected a population expansion event) (Table 5 and Online Resource 3). The ANT mismatch distribution observed curve did not perfectly fit the expected curve (the Harpending's RI was significant). However, in all other cases, tests of goodness of fit were not significant, and therefore, the observed curves fitted a population expansion model, as expected.

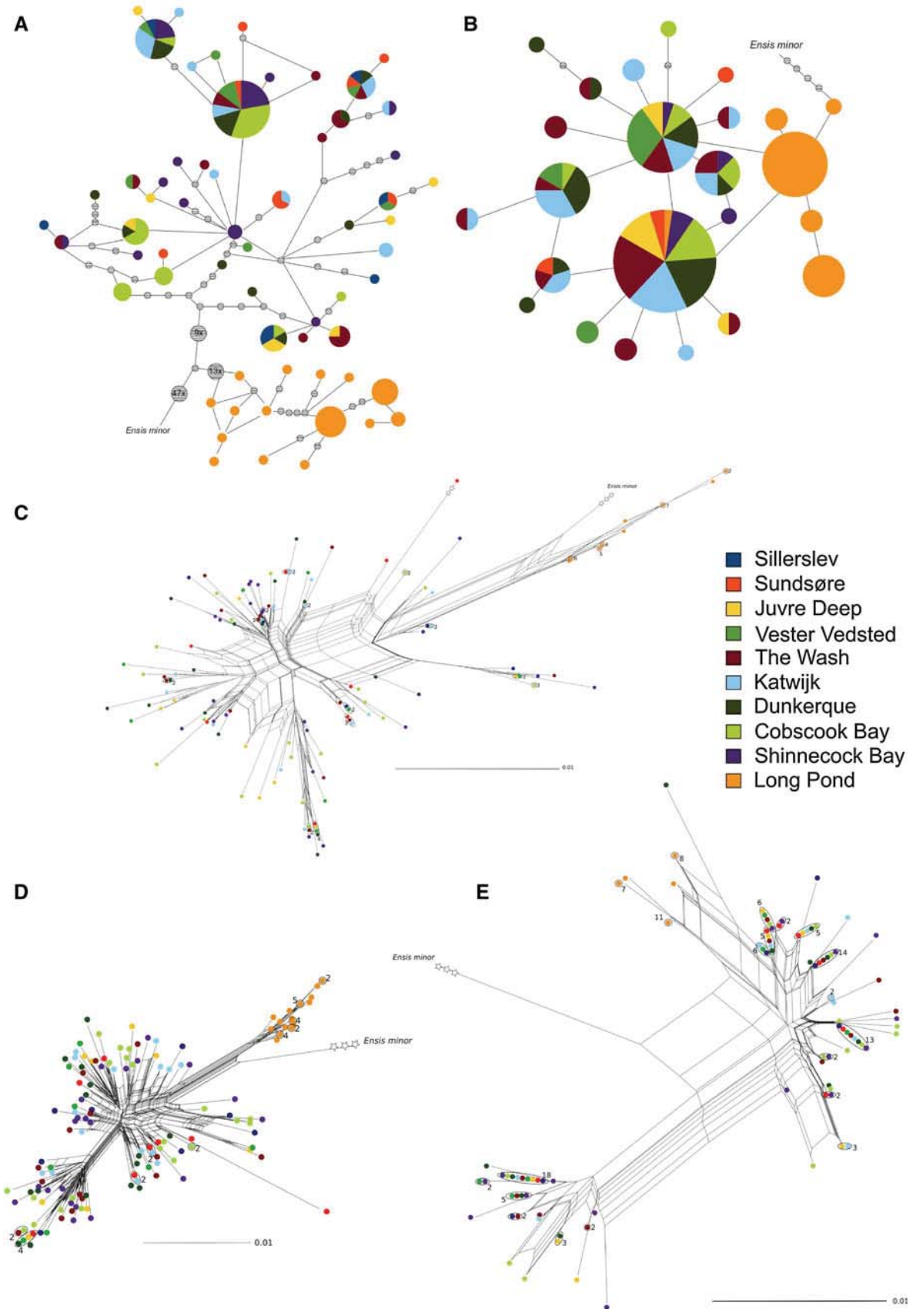
Estimates of time elapsed since expansion according to COI sequences were quite high, regardless of the mutation rate used. Expansion of the (hypothetical) source population of European individuals was, at the latest, 6 mya (Pleistocene), assuming all European individuals came from the same population and that there were no bottleneck effects during introduction or mass mortality events (Table 5). We should be cautious in relation to these results because the mutation rates employed could not be suitable for this species.

Fig. 3 Networks showing the phylogenetic relationships of obtained variants. **a** COI network. **b** ANT network. **c** ITS1 network. **d** ITS1 + ITS2 network. **E** ITS2 network. Each colour represents a sampling site, according to the legend. **a**, **b** Parsimony networks; the size of *circles* is proportional to variants frequency; *lined circles* are non-sampled variants inferred by the software; each line between circles represents a mutational step. **c**–**e** Distance networks; sequence-types with frequency = 1 are represented by *single dots*; sequence-types with higher frequencies are represented by more than one dot surrounded by ovals or circumferences out of which the absolute frequency is indicated by a *number*; lines between sequence-types are proportional to genetic distance; *lines* that have been shortened to fit in the figure were marked with a *star*

Evolutionary relationships among variants

The network obtained from the COI haplotype dataset (Fig. 3a) revealed two main haplogroups; one including all variants from Europe, Shinnecock Bay, and Cobscook Bay; and another one, separated by 28 mutational steps, which included all Long Pond haplotypes. Within the first haplogroup, no clear subgroups were recognised in terms of geography, that is, haplotypes from US and European sites were intermixed in the network. The ANT network (Fig. 3b) supported the separation of US/Europe and Long Pond individuals though not as clearly. In this case, all Canadian variants except one clustered apart all others. The remaining sequence-type corresponded to the one sampled in all sites except Vester Vedsted, which was the most frequent variant of the dataset. Networks performed with the ITS1, ITS2, and ITS1 + ITS2 variants (Fig. 3c, e) showed a similar picture. Long Pond sequence-types appeared clustered apart from all others in all three networks, and no subdivision by sampling site or geographic area was found among the remaining sequences (from Europe or US). Remarkably, in the case of the ITS1 + ITS2 concatenated dataset network (Fig. 3d), it was particularly evident that variants from the US and Europe were clearly distributed into two subgroups. This revealed the existence of two different ITS1-ITS2 (paralog) regions in the genomes of these animals, which were not present in the genomes of Long Pond individuals.

In the phylogenetic trees obtained from the COI + ANT + ITS1 + ITS2 concatenated dataset (Fig. 4), two reciprocally monophyletic groups were recognised, regardless of the phylogenetic method employed, and with the highest support. One group included all European sequences, those from Shinnecock Bay, and those from Cobscook Bay; and the other one, comprised all Long Pond sequences. K2P distances between those groups were: COI, 0.063 ± 0.012; ITS1, 0.028 ± 0.007; ITS2, 0.019 ± 0.006; and ANT, 0.005 ± 0.003.


A
B
Ensis minor
C
Ensis minor
Sillerslev
Sundsøre
Juvre Deep
Vester Vedsted
The Wash
Katwijk
Dunkerque
Cobscook Bay
Shinnecock Bay
Long Pond
D
Ensis minor
0.01
E
Ensis minor
0.01

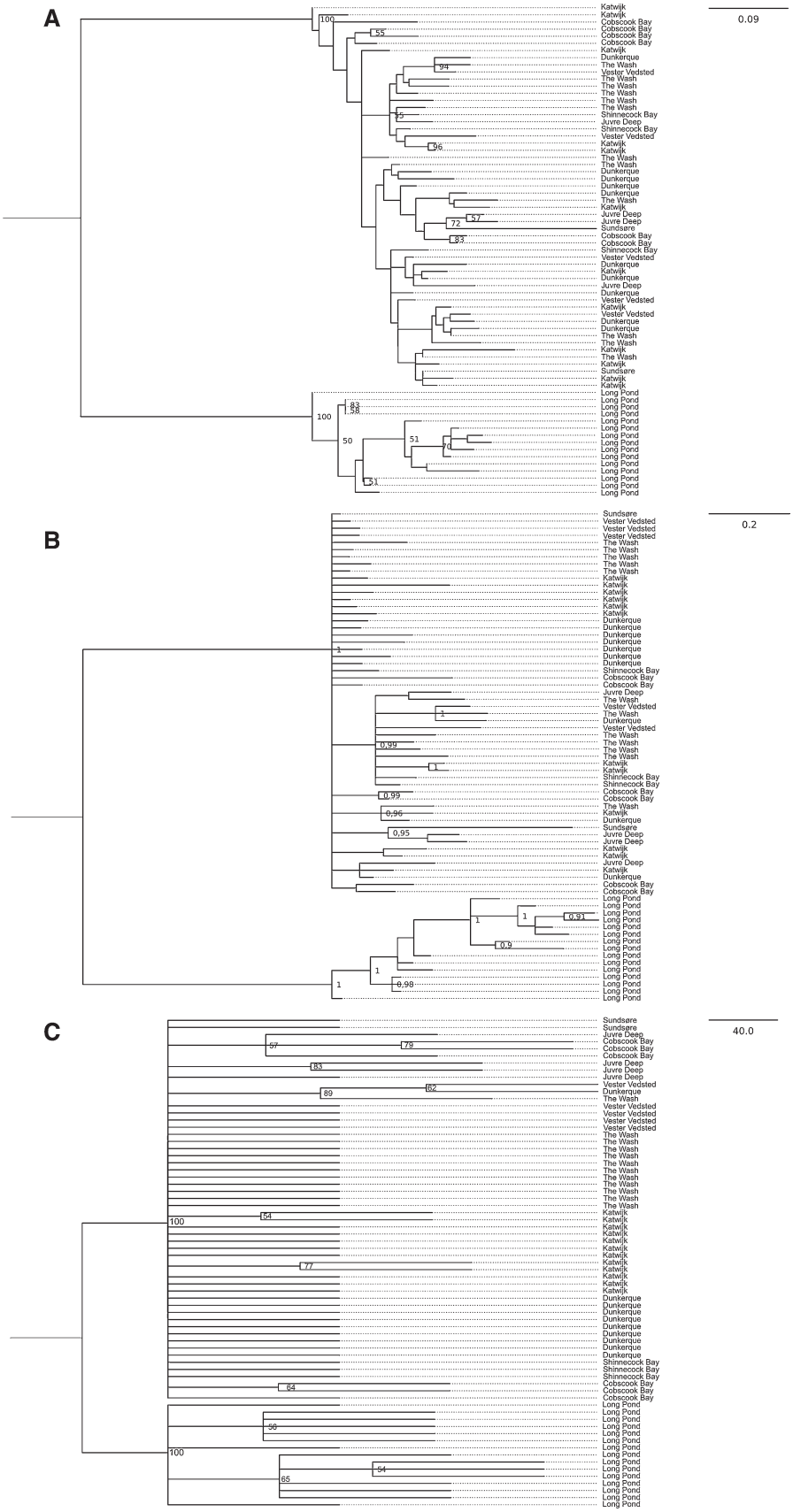


Fig. 4 Trees showing the phylogenetic relationships of sequences of the COI + ANT + ITS1 + ITS2 concatenated dataset. **a** Best-scoring maximum-likelihood tree (tree likelihood, −4,461.9); bootstrap values ≥50 indicated at the nodes. **b** Bayesian tree; posterior probability values ≥0.95 indicated at the nodes. **c** Maximum parsimony bootstrap 50 % majority-rule consensus tree; bootstrap values ≥50 indicated at the nodes

Shell morphometrics

All measurements obtained and the ratios calculated from them were recorded in Online Resource 4. Remarkably, 'distance b' values were greater for *Ensis* specimens from Long Pond than from Shinnecock Bay and Cobscook Bay. There was a significant site effect ($F_{2,\ 62} = 161.854$, p value = 0, ANCOVA; homoscedasticity among residuals according to Levenes's test and equal regression slopes). For example, for a shell of 135 mm, the mean 'distance b' value (95 % confidence interval in brackets) was 16.36 mm (15.65–17.08) in Long Pond, 7.47 mm (6.67–8.27) in Shinnecock Bay, and 9.01 mm (8.13–9.89) in Cobscook Bay. In Figs. 5 and 6, the relationships between 'distance b' values, and length or width were recorded per sampling site. To make an easy identification, a Long Pond individual showed a shell width (measured at the posterior adductor scar) less than twice 'distance b', whereas a shell from Shinnecock Bay or Cobscook Bay showed a width more than twice 'distance b'.

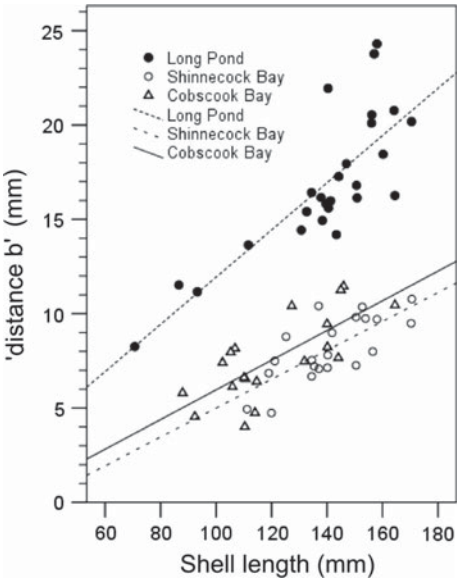


Fig. 5 The relationship of the distance along the dorsal line between the posterior end of the posterior adductor scar and the beginning of the pallial sinus ('*distance b*') and shell length of Ensis specimens. The coefficient of determination (r^2) for the linear relationship is 0.66, 0.54, and 0.56 for Long Pond, Shinnecock Bay and Cobscook Bay, respectively

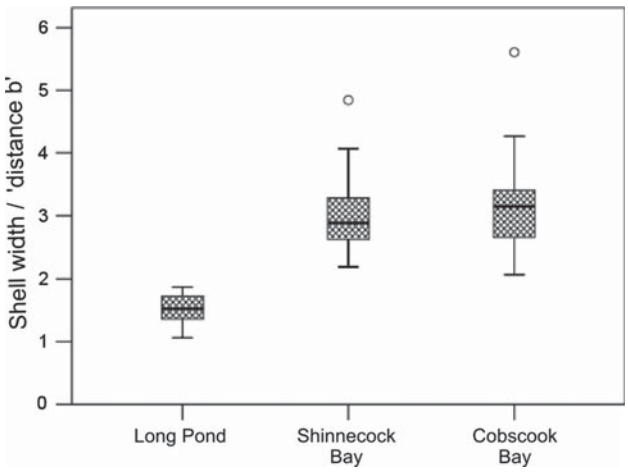


Fig. 6 Boxplot of the shell width—'*distance b*' ratio of *Ensis* specimens from Long Pond, Shinnecock Bay, and Cobscook Bay. '*Distance b*' is the distance along the dorsal line between the posterior end of the posterior adductor scar and the beginning of the pallial sinus

Taxonomy

Considering the genetic and morphometric results obtained in the present work, we propose to include the individuals from Long Pond within the new taxon *Ensis terranovensis* Vierna and Martínez-Lage sp.n. (Fig. 2, Online Resource 5).

Type material: MNCN-15.07/15001 (holotype; Fig. 2, Online Resource 5); MNCN-15.07/15002 (paratypes); MNHN-IM-2009-16706 to 16709 (paratypes); ZMUC-BIV-394 (paratypes); CMNML-096168 (paratypes). Type locality: Long Pond, Conception Bay, Newfoundland (Canada), 47° 30′ 54″ N, 52° 58′ 30″W. Individuals collected by scuba divers P. Sargent and R. O'Donell in 2007.

Etymology

The word *terranovensis* means 'from Newfoundland', referring to the area where the species was found.

Morphological description

In studied individuals ($n = 26$), shell length between 70.7 and 170.4 mm (mean 139.2 mm; SD 24.3 mm); shell width at the posterior adductor scar between 14.5 and 30.5 mm (mean 24.9 mm; SD 3.9 mm). Shell variable in curvature, in two individuals (whose lengths were 93.2 and 170.4 mm), straight or almost straight; in all others, curved to very curved. Valve margins parallel or almost parallel; some, slightly tapering posteriorly. Length/width ratio ranging between 4.8 and 7.2 (mean 5.6; SD 0.5). Anterior margin rounded, posterior margin truncated. Dorsal margin concave in some individuals, and almost or completely straight in others; ventral margin weakly to conspicuously convex.

Shell thick and strong (increasing with age); interior part, white coloured and slightly translucent. Posterior end of anterior adductor scar and posterior end of hinge, located at the same cross section, or more often, posterior

end of anterior adductor scar, slightly surpassing posterior end of hinge towards posterior margin. Anterior adductor scar at the anterior end, narrow, broadening posteriorly; its posterior end, usually rounded, but sometimes slightly truncated. Shape of posterior adductor scar, irregularly oval. The ratio 'distance b'/'distance a', ranging between 1.1 and 2.3 (mean 1.6; SD 0.3), therefore, distance between the posterior end of the posterior adductor scar and the beginning of the pallial sinus ('distance b'), always higher than the length of the posterior adductor scar itself ('distance a'). Dorsal pallial line, much closer to dorsal margin than ventral pallial line to ventral margin. Dorsal and ventral pallial lines not completely parallel to corresponding margins, quite irregular in some individuals. Pallial sinus formed by two concave areas separated by a convex one; dorsal part of pallial sinus more concave than ventral part; in general, pallial sinus shape, similar to an irregular W.

Exterior part of the shell with irregular growth lines; conspicuous and rough in older individuals. Younger individuals, olive green, dark brown, and grey coloured; older individuals, dark brown, whitish, and in some cases, reddish.

Many individuals displaying a bulge on the most anterior hinge tooth of the left valve.

Biotope and distribution

Razor shells were collected along a channel to a harbour (Long Pond), in Conception Bay, Newfoundland, Canada, from a depth of 6–11 m. The substrate consisted of a combination of hard to loose packed mud, sand, and ground sea shells. Animals were burrowed in this material 15–30 cm. In shallower waters (6 m), substrate turned to mostly fine sand. Since the sampling of additional sites within Newfoundland was out of the scope of this work, the distributional range of the species, apart from its occurrence in Conception Bay, remains unknown.

Remarks

In terms of genetics, our results indicate *E. terranovensis* and *E. directus* are two significantly different lineages regardless of the phylogenetic method employed, according to four different nuclear and mitochondrial markers. The sequences obtained here will be useful to complement morphological identifications of newly sampled individuals. In addition, the occurrence of two paralog ITS groups in *E. directus* and not in *E. terranovensis* is a further evidence supporting strong lineage divergence.

In terms of shell morphometrics, individuals from *E. terranovensis* are easily discernible from the other *Ensis* spp., including its sister taxon *E. directus*, after a careful analysis of the valves. *E. terranovensis* is characterised by having thicker and stronger valves comparing to *E. directus*, and individuals studied appeared to have undergone slower growth. The position of the posterior adductor scar is a clear and statistically significant difference between taxa; *E. terranovensis* individuals showed a shell width less than twice the distance between pallial sinus and posterior adductor scar, whereas a *E. directus* shell from Shinnecock Bay or Cobscook Bay showed a width more than twice that distance.

It should be pointed out that one individual from Graye–sur–Mer, Calvados, France (therefore, identified as *E. directus*) was reported to have a similar distance between pallial sinus and posterior adductor scar to *E. terranovensis* (Cosel 2009, Fig. 1f). No more data about this specimen are available, and its origin remains unclear.

Discussion

Genetic variation of *E. directus* in native and introduced sites

The introduction of *E. directus* in Europe is well documented. Several ecological studies focusing on this species have been conducted in the last decades (Swennen et al. 1985; Beukema and Dekker 1995; Armonies and Reise 1999; Cadée 2000; Palmer 2004; Krakau et al. 2006; Cardoso et al. 2009; Freudendahl et al. 2010; Tulp et al. 2010; Dannheim and Rumohr 2011; Dekker and Beukema 2012), but so far there have been no reports about the genetic variation of native and introduced populations except our report on ITS sequence variation of *Ensis* species (Vierna et al. 2010). After having recognised that the *Ensis* specimens from Long Pond belong to a new species, the conclusion about the applicability of ITS to differentiate among individuals from different geographic areas cannot be maintained.

One of our goals was to obtain preliminary information on the origin of European individuals. However, to identify the geographic source of introduced populations, the native range of the species must be thoroughly sampled and potential source populations must be sufficiently differentiated (Fitzpatrick et al. 2012). This is an important issue since the detection of source populations will depend not only on sampling intensity but also on the differentiation of these populations along the native range of the species. Indeed, our results suggest that both Shinnecock Bay and Cobscook Bay could be the origin of European individuals, since all of them could be assigned with high probability to one of these US sites. The site of Long Pond was completely discarded as potential source population. Nonetheless, if data from a third (southern) potential source

population become available, these results could be re-interpreted. Despite several attempts, we failed to obtain samples from the southern native range of the species that otherwise might have contributed to a more complete description of the genetic variation in the native range. Armonies and Reise (1999) discussed the possibility of linking the mass mortality events with the origin of European individuals. According to them, these individuals may have originated from an American population at the southern limit of its distributional range that may not be adapted to cold winter conditions supposed to cause mass mortality in Europe. However, it is unknown either to what extent native populations are subjected to mass mortality, and the role that factors other than storms may play in the mass mortality events, both in the native and in the introduced ranges. The higher differentiation of the European population to native ones compared to differentiation between Shinnecock Bay and Cobscook Bay might support that the source of European individuals is at the southern limit of the native range.

In their review on the genetic variation of various taxa in native and introduced populations, Dlugosch and Parker (2008) concluded that genetic variation is usually lower in introduced areas. However, according to Holland (2000), if the introduction involves a large (more than 1,000), genetically diverse assortment of individuals, we might expect to see little or no reduction in heterozygosity and allelic diversity relative to the gene pool of the source population. Here, after having sampled the main parts of the species' native and introduced ranges, we show that the European population as a whole, far from displaying low levels of genetic variation, is more variable than native sites in the US. In the same way, several sampling sites within the European range are more variable than native sites. For instance, Dunkerque is almost twice as variable as Cobscook Bay in terms of COI nucleotide diversity. The sites from the Limfjorden area are even more variable.

Taking into account the intensity of transatlantic shipping, the hypothesis of multiple-introduction events seems rather likely. The higher probability of Limfjorden, Danish Wadden Sea, and Katwijk to be assigned to Cobscook Bay seems to support this hypothesis. Multiple introductions appear to be common and contribute to increase variation (Dlugosch and Parker 2008). In this sense, the higher levels of variation detected in Europe agree with a multiple-introduction scenario. On the contrary, the time-line for the appearance of *E. directus* along the European coast suggests that colonisation happened gradually (for colonisation pattern see Cosel 2009). The very recent findings of isolated populations in western Britain (Paul Dansey personal communication; Dansey 2011) and northern Spain (Arias and Anadón 2012) could be a result of dispersion from European populations but it might also be a result of new introductions from native sites.

Apart from the already mentioned difficulties in determining the precise source population(s) and assessing the number of introductions, there is another factor that should be considered to understand the observed genetic variation both at native and introduced sites. The genetic variation of potential source populations that we have measured might not be the same as it was when introduction took place. In fact, temporal fluctuations in genetic variation have been reported for other *Ensis* (Varela et al. 2009, 2011). Even though it would have been ideal to have analysed individuals sampled in the native populations around 1978 (when the first migrants arrived into European waters), such samples are, to our best knowledge, unavailable. In the same way, fluctuations of genetic variation in the introduced range are also possible. In fact, Hedgecock and Pudovkin (2011) stressed the importance of studying temporal stability of genetic variation when conducting marine population genetic studies, but this is rarely done due to considerable time and economic limitations.

Therefore, we can conclude that (1) individuals from European sites can potentially be assigned to either Shinnecock Bay and Cobscook Bay but not to Long Pond. (2) Multiple introductions seem likely, but our data cannot prove this. And (3) potential bottlenecks during introduction(s) and mass mortality events seem not to have affected genetic variation in the introduced range.

Within the native range, the northern site (Cobscook Bay) is much less variable than Shinnecock Bay according to mitochondrial DNA, but not according to nuclear DNA, a fact that is quite intriguing. It is usual that northern populations in the Northern Hemisphere display less neutral genetic variation due to the effect of Quaternary glaciations (even though in marine animals, there are several reports of populations surviving in northern periglacial refugia, see Maggs et al. 2008; Krakau et al. 2012). However, nuclear and mitochondrial loci should be concordant. Therefore, a mitochondrial selective sweep in Cobscook Bay might be the cause of the reduced genetic variation. But neutrality tests did not detect deviations from neutrality so this question remains unanswered.

Our results suggest that F_{ST} is more conservative than tests of population differentiation to detect genetic differences among sites. Both Shinnecock Bay and Cobscook Bay show little but significant differentiation from Europe (considered as a representative of the source population) according to COI in the F_{ST} analyses, and to ITS1 in the test of population differentiation. Since the only significant comparison obtained between Shinnecock Bay and Cobscook Bay sites was the COI test of population differentiation, they seem to be less differentiated to each other than to Europe.

An absence of population structure over hundreds of kilometres of coast is not unexpected in populations of marine bivalves (e.g. Strasser and Barber 2008; Baker et al. 2008; Arnaud-Haond et al. 2008) since they are often characterised by frequent gene flow among sites. Nonetheless, there are also several examples of structure (e.g. Arnaud-Haond et al. 2008; Xiao et al. 2010; Mao et al. 2011). Dispersal and therefore gene flow is facilitated by external fertilisation and a planktonic larval stage but could also be reduced by marine currents and other physical barriers. The Western Atlantic Boreal Region extends from the Strait of Belle Isle to Cape Hatteras (Briggs and Bowen 2012). In this region, some main barriers to gene flow have been described, namely the Laurentian Channel, Cape Cod, and Cape Hatteras itself (see Fig. 1). The weak but significant population differentiation that we detected between Shinnecock Bay and Cobscook Bay may support Cape Cod as a barrier to gene flow.

The four sequence-based markers selected in this work showed different degrees of conservation, but all of them were informative both at the population and species levels. Indeed, the degree of conservation is expected to vary among different genomic regions. For example, in phylogenetics, more conserved genes are usually employed to resolve internal nodes in the phylogeny, whereas those more variable are, in general, able to resolve the external nodes. Here, we have shown that COI and ITS1 regions were more polymorphic than ITS2 and ANT. Population genetic results based on these four markers were, in general, concordant. Even though ITS2 and ANT sequences did not show extreme sequence differences among *E. directus* and *E. terranovensis* individuals (as COI and ITS1 did), they still separated both lineages, with the exception of the shared ANT sequence-type, a phenomenon known as incomplete lineage sorting. Because of this phenomenon, some haplotypes can remain identical in two isolated gene pools, a situation that mainly occurs when divergence is recent (April et al. 2011).

A new *Ensis* species

Taxonomy and systematics of *Ensis* razor shells have traditionally been based on shell morphology such as continuous shell characters (e.g. valve shape, length and shape of muscle scars, shell colour) (see Cosel 2009 and references therein), which are often overlapping among species. This absence of autapomorphies makes *Ensis* spp. a good candidate for combined studies of genes and morphometrics.

Though several species concepts have been proposed since the 1940s there is a working definition that considers species as separately evolving lineages. Kevin de Queiroz (2007) has proposed a unified species concept that treats this property as the only necessary property of species. Here, we demonstrate that *directus* and *terranovensis* are separately evolving lineages. There is a high genetic divergence between individuals belonging to *E. terranovensis* (the Long Pond population) and individuals belonging to *E. directus* (European, Shinnecock Bay, and Cobscook Bay sites) according to both nuclear and mitochondrial DNA that can only be explained by a two-lineages scenario. This divergence was confirmed by the morphometric analysis carried out.

In a recent paper, Kong et al. (2012) studied COI and ITS1 sequence data of the marine bivalves *Macrodiscus* spp. They described new species based on these two molecular markers and morphometrics, supporting the suitability of combining morphometrics and DNA to clarify taxonomy. K2P distances between *E. directus* and *E. terranovensis* were intermediate compared to the ones obtained for *Macrodiscus* spp.

The Laurentian Channel (Fig. 1) seems important in the speciation process of *E. directus* and *E. terranovensis* as it is a main barrier to gene flow between Coobscook Bay and Long Pond. This channel (>400 m depth) is a geographic barrier separating the Scotian Shelf and Newfoundland Shelf marine ecosystems (Sargent et al. 2008 and references therein) and seems to be a good candidate for vicariance.

Baker et al. (2008) detected significant (weak) population differentiation of the bivalve *Mercenaria mercenaria* between a population from Prince Edward Island (Gulf of St. Lawrence, Canada) and New York (US), and Hare and Weinberg (2005) found significant and strong genetic differentiation between a Îles de la Madeleine (Gulf of St. Lawrence) population of the bivalve *Spisula solidissima solidissima* and other Atlantic US populations. In the same way, Kenchington et al. (2006) detected significant genetic differentiation between Canadian populations of *Placopecten megallanicus* (at both sides of the Laurentian Channel), and Atlantic US populations. These examples support a genetic isolation of Gulf of St. Lawrence populations in bivalve species that might have facilitated speciation in *Ensis*. Nonetheless, it is unknown whether the two species co-occur somewhere in the NW Atlantic region. So, further studies on the present distribution of *E. directus* in its northernmost range and that of *E. terranovensis* are needed to unveil the evolutionary history of both taxa.

Conclusion

We found genetic variation at mitochondrial and nuclear markers in *E. directus* in its introduced (European) range to be higher than in native (North American) populations. This contrasts with our initial expectation of a strong effect

of random genetic drift due to both potential bottlenecks during introduction(s) and to mass mortalities observed several times in the introduced range. The observed patterns of genetic variation could be due to temporal fluctuations of genetic variation; to the fact that potential bottlenecks and mass mortalities might not have affected genetic variation in the introduced range; and to the occurrence of multiple introductions. The hypothesis of multiple introductions seems likely since it is supported by trained clustering analysis and by the intensity of transatlantic shipping. However, it contrasts with the observed gradual colonisation of European coastal waters. Population genetic analyses enabled us to identify a very divergent population from Newfoundland (Canada). Based on genetic and morphometric evidence, the examined specimens from this population belong to a new *Ensis* species that we described and named *E. terranovensis*.

Acknowledgments We are very grateful to Rudo von Cosel for his support and his drawing of *E. terranovensis* holotype, and to the following colleagues who helped us in some way or another during the execution of this work: André Martel, Anja Schulze, Barbara Buge, David Palmer, Diego Fonataneto, Ferruccio Maltagliati, Horacio Naveira, Jean-Marc Gagnon, Jean-Marie Dewarumez, Jeroen Goud, Jukka Corander, Manuel Pimentel, Mark Graham, Marta Vila, Neus Marí, Nicolas Puillandre, Ole S. Tendal, Paul Dansey, Philip Sargent, Rafael Araújo, Ray J. Thompson, Robert O'Donnell, Stephen T. Tettelbach, Tim Sheehan, and Tom Schioette. We would also like to thank two anonymous reviewers for critical and helpful comments. JV has been supported by a 'María Barbeito' fellowship and a travel grant, both from the Consellería de Economía e Industria, Xunta de Galicia (Spain) and the European Social Fund.

References

April J, Mayden RL, Hanner RH, Bernatchez L (2011) Genetic calibration of species diversity among North America's freshwater fishes. PNAS 108:10602–10607

Arias A, Anadón N (2012) First record of *Mercenaria mercenaria* (Bivalvia: Veneridae) and *Ensis directus* (Bivalvia: Pharidae) on Bay of Biscay, Iberian Peninsula. J Shellfish Res 31:57–60

Armonies W (2001) What an introduced species can tell us about the spatial extension of benthic populations. Mar Ecol Prog Ser 209:289–294

Armonies W, Reise K (1999) On the population development of the introduced razor clam *Ensis americanus* near the island pf Sylt (North Sea). Helgolan Mar Res 52:291–300

Arnaud-Haond S, Vonau V, Rouxel C, Bonhomme F, Prou J, Goyard E, Boudry P (2008) Genetic structure at different spatial scales in the pearl oyster (*Pinctada margaritifera cumingii*) in French Polynesian lagoons: beware of sampling strategy and genetic patchiness. Mar Biol 155:147–157

Audzijonyte A, Vrijenhoek RC (2010) Three nuclear genes for phylogenetic, SNP and population genetic studies of molluscs and other invertebrates. Mol Ecol Resour 10:200–204

Baker P, Austin JD, Bowen BW, Baker SM (2008) Range-wide population structure and history of the northern quahog (*Mercenaria merceneria*) inferred from mitochondrial DNA sequence data. ICES J Mar Sci 65:155–163

Beaumont MA (2007) Conservation genetics. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of statistical genetics, 3rd edn. Wiley, Chichester, pp 1021–1066

Beukema JJ, Dekker R (1995) Dynamics and growth of a recent invader into European coastal waters: the American razor clam, *Ensis directus*. J Mar Biol Assess UK 75:351–362

Briggs JC, Bowen BW (2012) A realignment of marine biogeographic provinces with particular reference to fish distributions. J Biogeogr 39:12–30

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255–265

Cadée GC (2000) Herring gulls feeding on a recent invader in the Wadden Sea, *Ensis directus*. In: Harper EM, Taylor JD, Crame JA (eds), *The evolutionary biology of the Bivalvia*. Geol Soc Lond Special Publ 177:459–464

Cardoso JFMF, Witt JIJ, van der Veer HW (2009) Reproductive investment of the American razor clam *Ensis americanus* in the Dutch Wadden Sea. J Sea Res 62:295–298

Cheng L, Connor TR, Aanensen DM, Spratt BG, Corander J (2011) Bayesian semi-supervised classification of bacterial samples using MLST databases. BMC Bioinformatics 12:302. doi: 10.1186/1471-2105-12-302

Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. Mol Ecol 9:1657–1660

Corander J, Marttinen P, Mäntyniemi S (2006) Bayesian identification of stock mixtures from molecular marker data. Fis B-NOAA 104:550–558

Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9:539

Cosel von R (2009) The razor shells of the eastern Atlantic, part 2. Pharidae II: the genus *Ensis* Schumacher, 1817 (Bivalvia, Solenoidea). Basteria 73:1–48

Dannheim J, Rumohr H (2011) The fate of an immigrant: *Ensis directus* in the eastern German Bight. Helgol Mar Res. doi: 10.1007/s10152-011-0271-2

Dansey P (2011) *Ensis directus* (Conrad 1843) (Bivalvia: Solenoidea) found in Liverpool Bay (Sea area S24). J Conchol 40:679

de Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56:879–886

Dekker R, Beukema JJ (2012) Long-term dynamics and productivity of a successful invader: The first three decades of the bivalve *Ensis directus* in the western Wadden Sea. J Sea Res http://dx.doi.org/10.1016/j.seares.2012.04.004

Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol 17:431–449

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2011) Geneious v5.4. Available from http://www.geneious.com

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Res 10:564–567

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Fitzpatrick BM, Fordyce JA, Niemiller ML, Reynolds RG (2012) What can DNA tell us about biological invasions? Biol Invasions 14:245–253

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase

subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3:294–299
Freudendahl ASL, Nielsen MM, Jensen T, Jensen KT (2010) The introduced clam *Ensis americanus* in the Wadden Sea: field experiment on impact of bird predation and tidal level on survival and growth. Helgoland Mar Res 64:93–100
Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking, and background selection. Genetics 147:915–925
Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709
Giegerich R, Meyer F, Schleiermacher C (1996) GeneFisher—software support for the detection of postulated genes. Proc Int Conf Intell Syst Mol Biol 4:68–77
Goudet J, Raymond M, de Meeüs T, Rousset F (1996) Testing differentiation in diploid populations. Genetics 144:1933–1940
Griffiths-Jones S (2005) RALEE—RNA ALignment editor in Emacs. Bioinformatics 21:257–259
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321
Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98
Hare MP, Weinberg JR (2005) Phylogeography of surfclams, *Spisula solidissima*, in the western North Atlantic based on mitochondrial and nuclear DNA sequences. Mar Biol 146:707–716
Harpending HC (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol 66:591–600
Heath DD, Rawson PD, Hilbish TJ (1995) PCR-based nuclear markers identify alien blue mussel (*Mytilus* spp.) genotypes on the west coast of Canada. Can J Fish Aquat Sci 52:2621–2627
Hedgecock D, Pudovkin AI (2011) Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bull Mar Sci 87:971–1002
Holland BS (2000) Genetics of marine bioinvasions. Hydrobiologia 420:63–71
Hudson R, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159
Huelsenbeck JP, Ronquist F (2001) MRBAYES: bayesian inference of phylogenetic trees. Bioinformatics 17:754–755
Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267
Kenchington E, Duggan R, Riddell T (1998) Early life history characteristics of the razor clam (*Ensis directus*) and the moonsnails (*Euspira* spp.) with applications to fisheries and aquaculture. Can Tech Rep Fish Aquat Sci 2223:1–32
Kenchington EL, Patwary MU, Zouros E, Bird CJ (2006) Genetic differentiation in relation to marine landscape in a broadcast-spawning bivalve mollusc (*Placopecten magellanicus*). Mol Ecol 15:1781–1796
Kong L, Matsukuma A, Hayashi I, Takada Y, Li Q (2012) Taxonomy of *Macridiscus* species (Bivalvia: Veneridae) from the western Pacific: insight based on molecular evidence, with description of a new species. J Moll Stud 78:1–11
Krakau M, Thieltges DW, Reise K (2006) Native parasites adopt introduced bivalves of the North Sea. Biol Invasions 8:919–925
Krakau M, Jacobsen S, Jensen KT, Reise K (2012) The cockle *Cerastoderma edule* at northeast Atlantic shores: genetic signatures of glacial refugia. Mar Biol 159:221–230
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, López R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452
Luttikhuizen PC, Drent J, Baker AJ (2003) Disjunct distribution of highly diverged mitochondrial lineage clade and population subdivision in a marine bivalve with pelagic larval dispersal. Mol Ecol 12:2215–2229
Maggs CA, Castilho R, Foltz D, Henzler C, Jolly MT, Kelly J, Olsen J, Perez KE, Stam W, Väinölä R, Viard F, Wares J (2008) Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. Ecology 89:S108–S122
Maine Sea Grant (2012) Accessed at http://www.seagrant.umaine.edu/resources-for-shellfish-growers/species/razor-clam. On 01 Feb 2012
Mao Y, Gao T, Yanagimoto T, Xiao Y (2011) Molecular phylogeography of *Ruditapes philippinarum* in the northwestern Pacific Ocean based on COI gene. J Exp Mar Biol Ecol 407:171–181
Marine Stewardship Council (2012) Accessed at http://www.msc.org/track-a-fishery/in-assessment/north-east-atlantic/dfa-dutch-north-sea-ensis. On 01 Feb 2012
McDonald J, Kreitman M (1991) Adaptative protein evolution at *adh* locus in *Drosophila*. Nature 351:652–654
Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the gateway computing environments workshop (GCE), 14 November 2010, New Orleans, LA pp 1–8
Mühlenhardt-Siegel U, Dörjes J, von Cosel R (1983) Die amerikanische Schwertmuschel *Ensis directus* (Conrad) in der Deutschen Bucht: 2. Populationsdynamik. Senckenb Marit 15:93–110
Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583
Palmer DW (2004) Growth of the razor clam *Ensis directus*, an alien species in the Wash on the east coast of England. J Mar Biol Assess UK 84:1075–1076
Raymond M, Rousset F (1995) An exact test for population differentiation. Evolution 49:1280–1283
Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569
Sargent PS, Methven DA, Hooper RG, McKenzie CH (2008) A range extension of the Atlantic silverside, *Menidia menidia*, to coastal waters of southwestern Newfoundland. Can Field Nat 122:338–344
Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690
Stamatakis A, Hoover P, Rougemont J (2008) A fast bootstrapping algorithm for the RAxML web-servers. Syst Biol 57:758–771
Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462
Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
Strasser CA, Barber PH (2008) Limited genetic variation and structure in softshell clams (*Mya arenaria*) across their native and introduced range. Conser Genet 10:803–814
Swennen C, Leopold MF, Stock M (1985) Notes on growth and behaviour of the American razor clam *Ensis directus* in the Wadden Sea and the predation on it by birds. Helgolan Mar Res 39:255–261
Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version. Sinauer Associates, Sunderland, MA, USA

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Tulp I, Craeymeersch J, Leopold M, van Damme C, Fey F, Verdaat H (2010) The role of the invasive bivalve *Ensis directus* as food source for fish and birds in the Dutch coastal zone. Estuar Coast Shelf Sci 90:116–128

Varela MA, Martínez-Lage A, González-Tizón AM (2009) Temporal genetic variation of microsatellite markers in the razor clam *Ensis arcuatus* (Bivalvia: Pharidae). J Mar Biol Assess UK 89:1703–1707

Varela MA, Martínez-Lage A, González-Tizón AM (2011) Genetic heterogeneity in natural beds of the razor clam *Ensis siliqua* revealed by microsatellites. J Mar Biol Assess UK 92:171–177

Vierna J, Martínez-Lage A, González-Tizón AM (2010) Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. Genome 53:23–34

von Cosel R, Dörjes J, Mühlenhardt-Siegel U (1982) Die amerikanische Schwertmuschel *Ensis directus* (Conrad) in der Deutschen Bucht. I. Zoogegraphis und Taxonomie im Vergleich mit den einheimischen Schwertmuschel-Arten. Senckenbergiana marit 14:147–173

Xiao J, Cordes JF, Wang H, Guo X, Reece KS (2010) Population genetics of *Crassostrea ariakensis* in Asia inferred from microsatellite markers. Mar Biol 157:1767–1781

Online Resource 1. Distribution of genetic variants per sampling site. Pie charts are divided into sections, and each section denotes a genetic variant. Coloured sections are variants which were sampled in at least two different sites. Therefore, it is possible to see which variants were shared among sites. White sections are variants private to a particular sampling site. The size of sections is proportional to the frequency of the corresponding variant in a given pie chart.

Sampling site	COI	ITS1	ITS2	ITS	ANT
Sillerslev					
Sundsøre					
Juvre Deep					
Vester Vedsted					
The Wash					
Katwijk					
Dunkerque					
Cobscook Bay					
Shinnecock Bay					
Long Pond					

Online Resource 2. Output of BAPS 5.3 analyses.

RESULTS OF INDIVIDUAL LEVEL MIXTURE ANALYSIS:
Number of clustered individuals: 77
Number of groups in optimal partition: 2
Log(marginal likelihood) of optimal partition: -4436.1688

Best Partition:
Cluster 1: {1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 14, 16, 17, 18,
19, 20, 22, 23, 26, 28, 29, 30, 31, 32, 35, 37,
38, 40, 41, 43, 45, 48, 49, 50, 51, 53, 54, 55,
57, 58, 60, 62, 63, 64, 65, 67, 70, 72, 76}
Cluster 2: {4, 11, 13, 15, 21, 24, 25, 27, 33, 34, 36, 39,
42, 44, 46, 47, 52, 56, 59, 61, 66, 68, 69, 71,
73, 74, 75, 77}

Posterior probability of assignment into clusters:

ind	Cobscook Bay	Shinnecock Bay
Limfjorden		
1	0.999219	0.000781
2	0.984484	0.015516
3	0.999999	0.000001
4	0.000032	0.999968
5	0.600783	0.399217
6	0.999917	0.000083
7	1.000000	0.000000
8	1.000000	0.000000
9	0.999980	0.000020
10	1.000000	0.000000
11	0.012190	0.987810
12	0.999058	0.000942
13	0.000038	0.999962
14	0.873325	0.126675
15	0.000287	0.999713
16	0.998215	0.001785
Danish Wadden Sea		
17	0.620531	0.379469
18	0.999855	0.000145
19	0.999664	0.000336
20	0.999602	0.000398
21	0.000009	0.999991
22	0.999998	0.000002
23	0.999998	0.000002
24	0.008427	0.991573
25	0.260782	0.739218
26	0.998774	0.001226
27	0.000026	0.999974
28	0.999537	0.000463
29	0.999912	0.000088
30	0.999556	0.000444
31	0.995200	0.004800
32	0.999891	0.000109
The Wash		
33	0.005685	0.994315
34	0.010122	0.989878
35	0.980660	0.019340
36	0.002348	0.997652
37	0.999640	0.000360
38	0.997939	0.002061
39	0.083628	0.916372
40	0.999956	0.000044
41	0.993733	0.006267
42	0.007937	0.992063
43	1.000000	0.000000
44	0.000150	0.999850

45	0.920481	0.079519
46	0.000131	0.999869
47	0.000131	0.999869

Katwijk

48	0.996636	0.003364
49	0.999983	0.000017
50	0.999324	0.000676
51	0.998816	0.001184
52	0.000002	0.999998
53	0.999978	0.000022
54	0.998784	0.001216
55	1.000000	0.000000
56	0.000243	0.999757
57	0.999651	0.000349
58	1.000000	0.000000
59	0.000042	0.999958
60	0.971257	0.028743
61	0.035650	0.964350
62	0.837755	0.162245

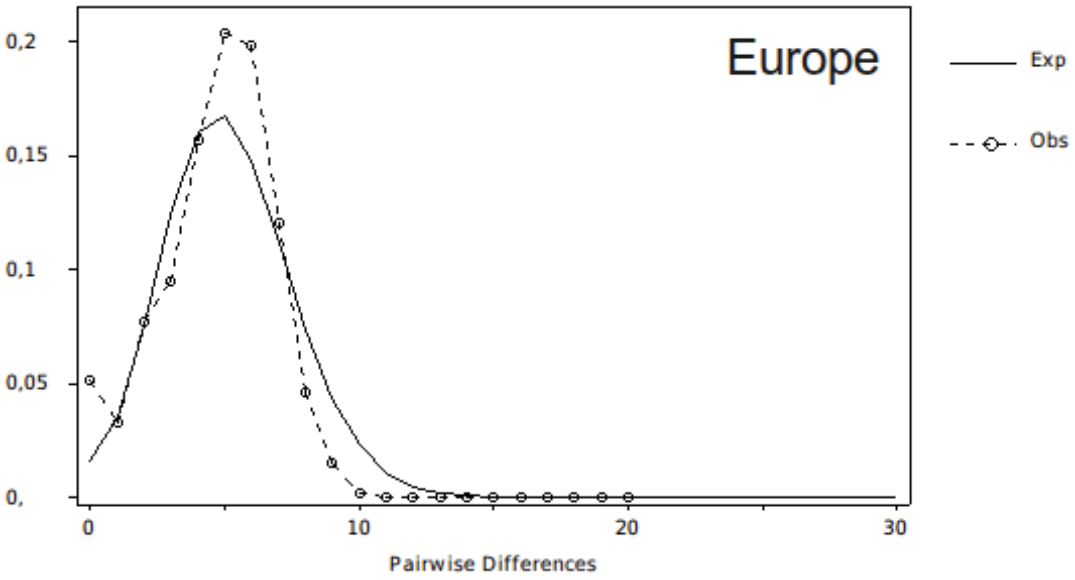
Dunkerque

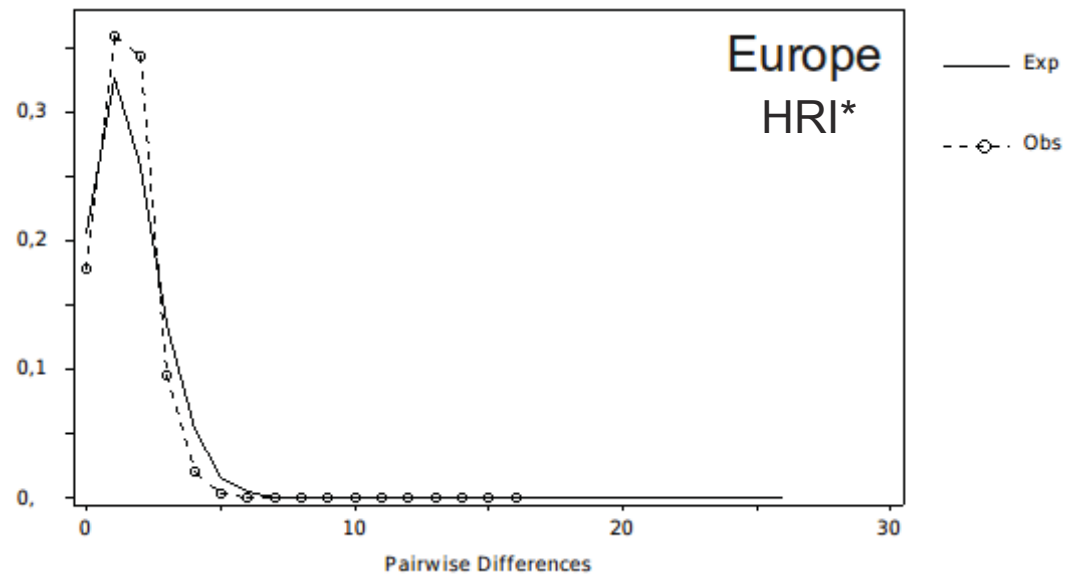
63	0.999990	0.000010
64	0.999933	0.000067
65	0.976061	0.023939
66	0.000008	0.999992
67	0.997629	0.002371
68	0.000054	0.999946
69	0.000478	0.999522
70	0.999991	0.000009
71	0.030563	0.969437
72	0.940740	0.059260
73	0.000009	0.999991
74	0.000058	0.999942
75	0.019122	0.980878
76	0.999876	0.000124
77	0.004508	0.995492

List of sizes of 10 best visited partitions and corresponding log(ml) values

2 -4436.168813
2 -4436.660616
2 -4440.728739
2 -4441.860201
2 -4441.900346
2 -4441.937434
2 -4443.309695
2 -4443.913968
2 -4444.324366
2 -4445.778860

Online Resource 3. Expected and observed mismatch distributions obtained from mitochondrial (COI) and single-copy nuclear (ANT) sequences of the European partition, under the "sudden expansion" model. COI, top; ANT, bottom. Goodness of fit of the observed and expected curves, assessed by the sum of squared deviations (SSD) and the raggedness statistic (α=0.05). Significance of these tests is indicated by an asterisk (*), when applicable.





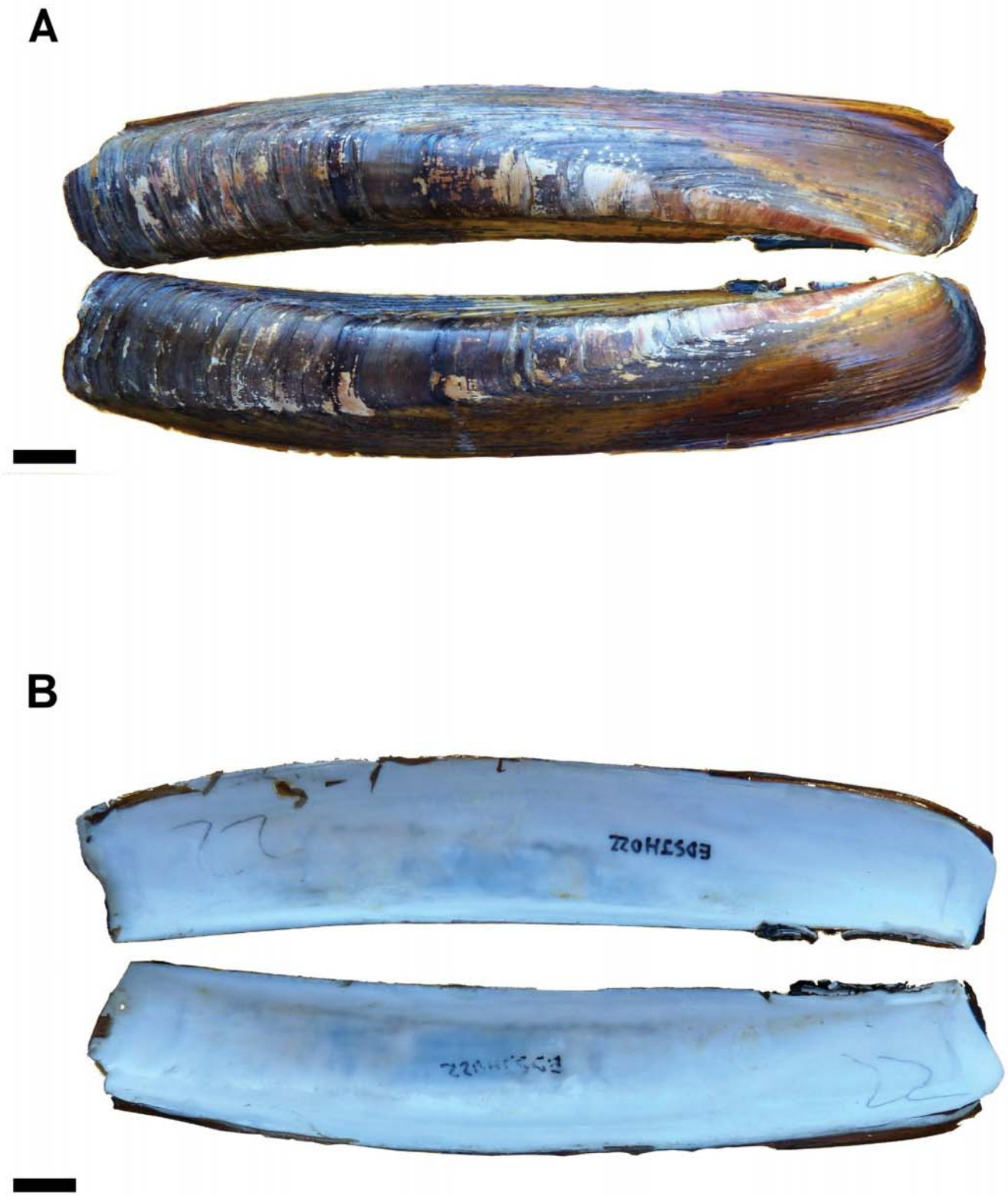
Online Resource 4. Shell morphometrics: Measurements obtained and ratios calculated from different morphological characteristics.

Individual	Bulge on hinge tooth?	Shell weight (g)	Length (mm)	Width (mm)	Distance a (mm)	Distance b (mm)	Length / Width	b/a	a / Width	a / Length	b / Width	b / Length
Long Pond 001	broken	19.582	156.000	27.360	10.640	20.100	5.702	1.889	0.389	0.068	0.735	0.129
Long Pond 002	no	18.720	157.000	25.300	10.240	23.770	6.206	2.321	0.405	0.065	0.940	0.151
Long Pond 003	yes	*20.308*	146.950	26.350	10.220	17.950	5.577	1.756	0.388	0.070	0.681	0.122
Long Pond 004	yes	6.769	111.600	20.490	10.050	13.640	5.447	1.357	0.490	0.090	0.666	0.122
Long Pond 005	yes	*22.346*	150.640	20.880	10.210	16.800	7.215	1.645	0.489	0.068	0.805	0.112
Long Pond 006	yes	*26.004*	138.310	27.890	12.290	14.940	4.959	1.216	0.441	0.089	0.536	0.108
Long Pond 007	yes	19.510	144.180	26.390	11.460	17.270	5.463	1.507	0.434	0.079	0.654	0.120
Long Pond 008	broken	20.305	132.640	27.800	10.370	15.410	4.771	1.486	0.373	0.078	0.554	0.116
Long Pond 009	yes	23.978	141.290	26.350	12.820	15.980	5.362	1.246	0.487	0.091	0.606	0.113
Long Pond 010	yes	16.045	140.510	24.460	11.330	15.600	5.744	1.377	0.463	0.081	0.638	0.111
Long Pond 011	no	18.480	140.360	26.410	10.320	21.940	5.315	2.126	0.391	0.074	0.831	0.156
Long Pond 012	no	*21.072*	143.350	25.830	11.080	14.190	5.550	1.281	0.429	0.077	0.549	0.099
Long Pond 013	broken	3.314	70.660	14.470	5.450	8.260	4.883	1.516	0.377	0.077	0.571	0.117
Long Pond 015	yes	12.888	130.700	22.590	10.570	14.430	5.786	1.365	0.468	0.081	0.639	0.110
Long Pond 016	yes	20.767	139.640	24.710	10.300	15.820	5.651	1.536	0.417	0.074	0.640	0.113
Long Pond 017	yes	15.652	134.400	22.660	12.080	16.420	5.931	1.359	0.533	0.090	0.725	0.122
Long Pond 018	broken	19.318	137.840	25.950	11.340	16.160	5.312	1.425	0.437	0.082	0.623	0.117
Long Pond 019	no	4.700	93.240	19.260	5.510	11.160	4.841	2.025	0.286	0.059	0.579	0.120
Long Pond 020	broken	3.799	86.560	16.880	6.640	11.520	5.128	1.735	0.393	0.077	0.682	0.133
Long Pond 021	yes	*24.034*	150.820	30.210	10.040	16.140	4.992	1.608	0.332	0.067	0.534	0.107
Long Pond 022	yes	24.055	160.200	28.010	9.940	18.450	5.719	1.856	0.355	0.062	0.659	0.115
Long Pond 024	maybe	19.242	158.000	25.840	11.320	24.300	6.115	2.147	0.438	0.072	0.940	0.154
Long Pond 025	yes	25.617	164.440	28.460	12.260	16.260	5.778	1.326	0.431	0.075	0.571	0.099
Long Pond 026	yes	26.941	164.150	26.440	10.250	20.770	6.208	2.026	0.388	0.062	0.786	0.127
Long Pond 027	yes	33.965	170.450	30.510	18.960	20.180	5.587	1.064	0.621	0.111	0.661	0.118
Long Pond 028	yes	25.411	156.100	27.080	11.590	20.540	5.764	1.772	0.428	0.074	0.758	0.132
Mean		18.955	139.232	24.945	10.665	16.846	5.577	1.614	0.426	0.077	0.676	0.121
Variance		56.328	590.983	15.012	6.240	14.092	0.276	0.108	0.004	0.000	0.013	0.000
Standard deviation		7.505	24.310	3.875	2.498	3.754	0.525	0.329	0.067	0.011	0.113	0.015
Shinnecock Bay 001	broken	*8.900*	121.140	22.800	7.050	7.490	5.313	0.941	0.309	0.058	0.329	0.062
Shinnecock Bay 002	broken	*15.382*	137.070	24.030	6.800	7.080	5.704	0.960	0.283	0.050	0.295	0.052
Shinnecock Bay 003	no	19.347	153.870	25.550	10.510	9.740	6.022	1.079	0.411	0.068	0.381	0.063
Shinnecock Bay 004	no	*8.308*	135.350	21.970	6.740	7.210	6.161	0.935	0.307	0.050	0.328	0.053
Shinnecock Bay 005	broken	10.031	119.960	22.960	7.790	4.740	5.225	1.643	0.339	0.065	0.206	0.040
Shinnecock Bay 006	broken	*4.852*	111.150	20.100	5.810	4.940	5.530	1.176	0.289	0.052	0.246	0.044
Shinnecock Bay 007	no	20.993	170.270	27.380	11.840	9.490	6.219	1.248	0.432	0.070	0.347	0.056
Shinnecock Bay 008	broken	6.363	119.000	18.200	6.700	6.850	6.538	0.978	0.368	0.056	0.376	0.058
Shinnecock Bay 009	no	16.366	150.430	25.200	8.860	9.820	5.969	0.902	0.352	0.059	0.390	0.065
Shinnecock Bay 010	no	24.658	170.540	26.800	10.380	10.780	6.363	0.963	0.387	0.061	0.402	0.063
Shinnecock Bay 011	no	12.021	134.500	24.740	9.440	7.530	5.437	1.254	0.382	0.070	0.304	0.056
Shinnecock Bay 012	no	9.875	134.430	22.430	7.470	6.680	5.993	1.118	0.333	0.056	0.298	0.050
Shinnecock Bay 013	no	10.591	140.140	20.980	7.860	7.140	6.680	1.101	0.375	0.056	0.340	0.051
Shinnecock Bay 014	broken	10.438	141.820	21.870	7.340	8.990	6.485	0.816	0.336	0.052	0.411	0.063
Shinnecock Bay 015	broken	11.418	140.260	23.510	10.310	7.800	5.966	1.322	0.439	0.074	0.332	0.056
Shinnecock Bay 016	broken	18.197	158.000	25.730	9.910	9.700	6.141	1.022	0.385	0.063	0.377	0.061
Shinnecock Bay 017	broken	19.612	152.680	26.420	9.940	10.360	5.779	0.959	0.376	0.065	0.392	0.068
Shinnecock Bay 018	broken	14.146	150.460	24.780	8.020	7.260	6.072	1.105	0.324	0.053	0.293	0.048
Shinnecock Bay 019	no	10.460	136.980	22.810	6.040	10.410	6.005	0.580	0.265	0.044	0.456	0.076
Shinnecock Bay 020	broken	18.488	156.500	20.960	7.040	7.990	7.467	0.881	0.336	0.045	0.381	0.051
Shinnecock Bay 021	broken	10.664	125.200	23.630	8.320	8.780	5.298	0.948	0.352	0.066	0.372	0.070
Cobscook Bay 001	maybe	31.018	164.470	30.410	11.180	10.450	5.408	0.935	0.368	0.068	0.344	0.064
Cobscook Bay 002	maybe	13.347	144.030	24.700	7.170	7.660	5.831	1.068	0.290	0.050	0.310	0.053
Cobscook Bay 003	no	21.629	145.960	29.690	11.460	11.450	4.916	0.999	0.386	0.079	0.386	0.078
Cobscook Bay 004	broken	*15.462*	144.940	25.380	10.370	11.260	5.711	1.086	0.409	0.072	0.444	0.078
Cobscook Bay 005	no	5.927	110.580	20.610	6.440	6.550	5.365	1.017	0.312	0.058	0.318	0.059
Cobscook Bay 006	no	9.518	110.050	24.220	7.100	6.610	4.544	0.931	0.293	0.065	0.273	0.060
Cobscook Bay 007	no	5.161	106.800	20.400	6.120	8.160	5.235	1.333	0.300	0.057	0.400	0.076
Cobscook Bay 009	no	*6.538*	105.880	22.840	7.030	6.130	4.636	0.872	0.308	0.066	0.268	0.058
Cobscook Bay 010	broken	*2.630*	87.910	15.740	4.630	5.800	5.585	1.253	0.294	0.053	0.368	0.066
Cobscook Bay 011	broken	5.438	114.010	20.280	7.650	4.750	5.622	0.621	0.377	0.067	0.234	0.042
Cobscook Bay 012	broken	2.936	92.320	15.170	3.510	4.540	6.086	1.293	0.231	0.038	0.299	0.049
Cobscook Bay 013	broken	5.134	105.260	20.200	7.250	7.960	5.211	1.098	0.359	0.069	0.394	0.076
Cobscook Bay 014	broken	14.924	140.140	28.500	10.560	8.230	4.917	0.779	0.371	0.075	0.289	0.059
Cobscook Bay 015	maybe	10.449	131.830	23.900	10.430	7.480	5.516	0.717	0.436	0.079	0.313	0.057
Cobscook Bay 016	broken	7.293	110.330	22.490	8.750	4.010	4.906	0.458	0.389	0.079	0.178	0.036
Cobscook Bay 017	maybe	8.525	127.280	21.540	11.450	10.410	5.909	0.909	0.532	0.090	0.483	0.082
Cobscook Bay 018	no	17.722	139.990	27.350	10.190	9.450	5.118	0.927	0.373	0.073	0.346	0.068
Cobscook Bay 020	maybe	*5.396*	102.300	20.370	9.570	7.410	5.022	0.774	0.470	0.094	0.364	0.072
Cobscook Bay 021	broken	7.416	114.620	21.440	6.140	6.410	5.346	1.044	0.286	0.054	0.299	0.056
Mean		11.939	131.461	23.202	8.279	7.888	5.681	1.001	0.354	0.063	0.339	0.060
Variance		40.576	441.055	10.923	4.081	3.810	0.356	0.049	0.004	0.000	0.004	0.000
Standard deviation		6.370	21.001	3.305	2.020	1.952	0.596	0.221	0.060	0.012	0.065	0.011
Shinnecock Bay 001	broken	*8.900*	121.140	22.800	7.050	7.490	5.313	0.941	0.309	0.058	0.329	0.062
Shinnecock Bay 002	broken	*15.382*	137.070	24.030	6.800	7.080	5.704	0.960	0.283	0.050	0.295	0.052
Shinnecock Bay 003	no	19.347	153.870	25.550	10.510	9.740	6.022	1.079	0.411	0.068	0.381	0.063
Shinnecock Bay 004	no	*8.308*	135.350	21.970	6.740	7.210	6.161	0.935	0.307	0.050	0.328	0.053
Shinnecock Bay 005	broken	10.031	119.960	22.960	7.790	4.740	5.225	1.643	0.339	0.065	0.206	0.040
Shinnecock Bay 006	broken	*4.852*	111.150	20.100	5.810	4.940	5.530	1.176	0.289	0.052	0.246	0.044
Shinnecock Bay 007	no	20.993	170.270	27.380	11.840	9.490	6.219	1.248	0.432	0.070	0.347	0.056
Shinnecock Bay 008	broken	6.363	119.000	18.200	6.700	6.850	6.538	0.978	0.368	0.056	0.376	0.058
Shinnecock Bay 009	no	16.366	150.430	25.200	8.860	9.820	5.969	0.902	0.352	0.059	0.390	0.065
Shinnecock Bay 010	no	24.658	170.540	26.800	10.380	10.780	6.363	0.963	0.387	0.061	0.402	0.063
Shinnecock Bay 011	no	12.021	134.500	24.740	9.440	7.530	5.437	1.254	0.382	0.070	0.304	0.056
Shinnecock Bay 012	no	9.875	134.430	22.430	7.470	6.680	5.993	1.118	0.333	0.056	0.298	0.050
Shinnecock Bay 013	no	10.591	140.140	20.980	7.860	7.140	6.680	1.101	0.375	0.056	0.340	0.051

Online Resource 4.

Individual	Bulge on hinge tooth?	Shell weight (g)	Length (mm)	Width (mm)	Distance a (mm)	Distance b (mm)	Length / Width	b/a	a / Width	a / Length	b / Width	b / Length
Shinnecock Bay 014	broken	10.438	141.820	21.870	7.340	8.990	6.485	0.816	0.336	0.052	0.411	0.063
Shinnecock Bay 015	broken	11.418	140.260	23.510	10.310	7.800	5.966	1.322	0.439	0.074	0.332	0.056
Shinnecock Bay 016	broken	18.197	158.000	25.730	9.910	9.700	6.141	1.022	0.385	0.063	0.377	0.061
Shinnecock Bay 017	broken	19.612	152.680	26.420	9.940	10.360	5.779	0.959	0.376	0.065	0.392	0.068
Shinnecock Bay 018	broken	14.146	150.460	24.780	8.020	7.260	6.072	1.105	0.324	0.053	0.293	0.048
Shinnecock Bay 019	no	10.460	136.980	22.810	6.040	10.410	6.005	0.580	0.265	0.044	0.456	0.076
Shinnecock Bay 020	broken	18.488	156.500	20.960	7.040	7.990	7.467	0.881	0.336	0.045	0.381	0.051
Shinnecock Bay 021	broken	10.664	125.200	23.630	8.320	8.780	5.298	0.948	0.352	0.066	0.372	0.070
Mean		5.401	55.329	9.034	3.187	3.395	2.343	0.363	0.134	0.022	0.144	0.024
Variance		16.005	125.505	3.482	2.426	1.400	0.396	0.047	0.003	0.000	0.002	0.000
Standard deviation		4.001	11.203	1.866	1.558	1.183	0.629	0.216	0.051	0.011	0.049	0.010
Cobscook Bay 001	maybe	31.018	164.470	30.410	11.180	10.450	5.408	0.935	0.368	0.068	0.344	0.064
Cobscook Bay 002	maybe	13.347	144.030	24.700	7.170	7.660	5.831	1.068	0.290	0.050	0.310	0.053
Cobscook Bay 003	no	21.629	145.960	29.690	11.460	11.450	4.916	0.999	0.386	0.079	0.386	0.078
Cobscook Bay 004	broken	*15.462*	144.940	25.380	10.370	11.260	5.711	1.086	0.409	0.072	0.444	0.078
Cobscook Bay 005	no	5.927	110.580	20.610	6.440	6.550	5.365	1.017	0.312	0.058	0.318	0.059
Cobscook Bay 006	no	9.518	110.050	24.220	7.100	6.610	4.544	0.931	0.293	0.065	0.273	0.060
Cobscook Bay 007	no	5.161	106.800	20.400	6.120	8.160	5.235	1.333	0.300	0.057	0.400	0.076
Cobscook Bay 009	no	*6.538*	105.880	22.840	7.030	6.130	4.636	0.872	0.308	0.066	0.268	0.058
Cobscook Bay 010	broken	*2.630*	87.910	15.740	4.630	5.800	5.585	1.253	0.294	0.053	0.368	0.066
Cobscook Bay 011	broken	5.438	114.010	20.280	7.650	4.750	5.622	0.621	0.377	0.067	0.234	0.042
Cobscook Bay 012	broken	2.936	92.320	15.170	3.510	4.540	6.086	1.293	0.231	0.038	0.299	0.049
Cobscook Bay 013	broken	5.134	105.260	20.200	7.250	7.960	5.211	1.098	0.359	0.069	0.394	0.076
Cobscook Bay 014	broken	14.924	140.140	28.500	10.560	8.230	4.917	0.779	0.371	0.075	0.289	0.059
Cobscook Bay 015	maybe	10.449	131.830	23.900	10.430	7.480	5.516	0.717	0.436	0.079	0.313	0.057
Cobscook Bay 016	broken	7.293	110.330	22.490	8.750	4.010	4.906	0.458	0.389	0.079	0.178	0.036
Cobscook Bay 017	maybe	8.525	127.280	21.540	11.450	10.410	5.909	0.909	0.532	0.090	0.483	0.082
Cobscook Bay 018	no	17.722	139.990	27.350	10.190	9.450	5.118	0.927	0.373	0.073	0.346	0.068
Cobscook Bay 020	maybe	*5.396*	102.300	20.370	9.570	7.410	5.022	0.774	0.470	0.094	0.364	0.072
Cobscook Bay 021	broken	7.416	114.620	21.440	6.140	6.410	5.346	1.044	0.286	0.054	0.299	0.056
Mean		10.340	120.984	22.907	8.263	7.617	5.310	0.953	0.357	0.068	0.332	0.063
Variance		52.123	437.606	17.390	5.702	4.879	0.182	0.050	0.005	0.000	0.005	0.000
Standard deviation		7.220	20.919	4.170	2.388	2.209	0.426	0.224	0.073	0.014	0.073	0.013
Long Pond	Mean	18.955	139.232	24.945	10.665	16.846	5.577	1.614	0.426	0.077	0.676	0.121
	Standard deviation	7.505	24.310	3.875	2.498	3.754	0.525	0.329	0.067	0.011	0.113	0.015
Shinnecock Bay and Cobscook Bay	Mean	11.939	131.461	23.202	8.279	7.888	5.681	1.001	0.354	0.063	0.339	0.060
	Standard deviation	6.370	21.001	3.305	2.020	1.952	0.596	0.221	0.060	0.012	0.065	0.011
Shinnecock Bay	Mean	13.386	140.940	23.469	8.294	8.132	6.017	1.044	0.351	0.059	0.346	0.057
	Standard deviation	5.252	16.358	2.344	1.681	1.705	0.529	0.215	0.047	0.008	0.059	0.009
Cobscook Bay	Mean	10.340	120.984	22.907	8.263	7.617	5.310	0.953	0.357	0.068	0.332	0.063
	Standard deviation	7.220	20.919	4.170	2.388	2.209	0.426	0.224	0.073	0.014	0.073	0.013

Online Resource 5. *Ensis terranovensis* n.sp. Vierna & Martínez-Lage, holotype. **A** Valves, external view. **B** Valves, internal view. **C** Right valve, hinge region. **D** Left valve, hinge region. **E** Right valve, pallial sinus. **F** Left valve, pallial sinus. Scale bar = 1 cm.



C

D

E

F

4.4 Species delimitation and DNA barcoding of Atlantic *Ensis* (Bivalvia, Pharidae)

Joaquín Vierna, Joël Cuperus, Andrés Martínez-Lage, Jeroen M. Jansen, Alejandra Perina, Hilde Van Pelt, Ana M. González-Tizón (2013) Species delimitation and DNA barcoding of Atlantic *Ensis* (Bivalvia, Pharidae). Zoologica Scripta doi:10.1111/zsc.12038.

Bibliometrics 2012 JCR Science Edition

Impact factor: 2.793

Evolutionary Biology: Q2

Zoology: Q1

Zoologica Scripta

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Species delimitation and DNA barcoding of Atlantic *Ensis* (Bivalvia, Pharidae)

Joaquín Vierna, Joël Cuperus, Andrés Martínez-Lage, Jeroen M. Jansen, Alejandra Perina, Hilde Van Pelt & Ana M. González-Tizón

Submitted: 17 May 2013
Accepted: 11 August 2013
doi:10.1111/zsc.12038

Vierna, J., Cuperus, J., Martínez-Lage, A., Jansen, J.M., Perina, A., Van Pelt, H. & González-Tizón, A.M. (2013). Species delimitation and DNA barcoding of Atlantic *Ensis* (Bivalvia, Pharidae). — *Zoologica Scripta*, *00*, 000–000.

Ensis Schumacher, 1817 razor shells occur at both sides of the Atlantic and along the Pacific coasts of tropical west America, Peru, and Chile. Many of them are marketed in various regions. However, the absence of clear autapomorphies in the shell and the sympatric distributions of some species often prevent a correct identification of specimens. As a consequence, populations cannot be properly managed, and edible species are almost always mislabelled along the production chain. In this work, we studied whether the currently accepted Atlantic *Ensis* morphospecies are different evolutionary lineages, to clarify their taxonomic status and enable molecular identifications through DNA barcoding. For this, we studied 109 specimens sampled at 27 sites, which were identified as belonging to nine of those morphospecies. We analysed nucleotide variation at four nuclear (18S, 5.8S, ITS1, and ITS2) and two mitochondrial (COI and 16S) regions, although the 18S and 5.8S regions were not informative at the species level and were not further considered. The phylogenetic trees and networks obtained supported all morphospecies as separately evolving lineages. Phylogenetic trees recovered *Ensis* at each side of the Atlantic as reciprocally monophyletic. Remarkably, we confirm the co-occurrence of the morphologically similar *E. minor* (Chenu, 1843) and *E. siliqua* (Linné, 1758) along the NW Iberian coast, a fact that has been often overlooked. In South America, a relevant divergence between *E. macha* (Molina, 1792) individuals from Chile and Argentina was unveiled and suggests incipient speciation. We also confirm the occurrence of the North American species *E. directus* (Conrad, 1843) as far south as north-eastern Florida. Among the genomic regions analysed, we suggest COI as the most suitable DNA barcode for Atlantic *Ensis*. Our results will contribute to the conservation and management of *Ensis* populations and will enable reliable identifications of the edible species, even in the absence of the valves. The name *Ensis coseli* Vierna nom. nov. is proposed to replace *E. minor* Dall, 1899 non (Chenu, 1843).

Corresponding author: *Joaquín Vierna, Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, A Fraga 10, A Coruña, E-15008, Spain. E-mails: jvierna@udc.es, joaquin@allgenetics.eu*

Joaquín Vierna, Evolutionary Biology Group (GIBE), Department of Molecular and Cell Biology, Universidade da Coruña, A Fraga 10, A Coruña, E-15008, Spain and AllGenetics, Edificio de Servicios Centrales de Investigación, Campus de Elviña s/n, A Coruña, E-15008, Spain

Joël Cuperus, IMARES Wageningen UR, Ambachtsweg 8a, Den Helder, NL-1785 AJ, The Netherlands

Andrés Martínez-Lage, Evolutionary Biology Group (GIBE), Department of Molecular and Cell Biology, Universidade da Coruña, A Fraga 10, A Coruña, E-15008, Spain

Jeroen M. Jansen, IMARES Wageningen UR, Ambachtsweg 8a, Den Helder, NL-1785 AJ, The Netherlands

Alejandra Perina, Evolutionary Biology Group (GIBE), Department of Molecular and Cell Biology, Universidade da Coruña, A Fraga 10, A Coruña, E-15008, Spain and AllGenetics, Edificio de Servicios Centrales de Investigación, Campus de Elviña s/n, A Coruña, E-15008, Spain

Hilde Van Pelt, IMARES Wageningen UR, Ambachtsweg 8a, Den Helder, NL-1785 AJ, The Netherlands

Ana M. González-Tizón, Evolutionary Biology Group (GIBE), Department of Molecular and Cell Biology, Universidade da Coruña, A Fraga 10, A Coruña, E-15008, Spain

Introduction

Ensis Schumacher, 1817 razor shells are a group of marine bivalve molluscs, characterised by their elongated shells, which are often found along the coasts at both sides of the Atlantic Ocean and along the Pacific coasts of tropical west America, Peru, and Chile. Several species are considered a delicacy and are marketed in some European, South American, and North American regions. For instance, in Chile alone, almost 6000 tons of *E. macha* (Molina, 1792) were landed in 1999 (Barón *et al.* 2004) and 1400 tons during 2005 (Ariz Abarca *et al.* 2007).

In the Atlantic, the genus is composed of 9–10 extant morphospecies that inhabit sandy and fine gravel substrata with limited exposure to wave action, from the intertidal to a depth of ca. 80 m (Cosel 2009). *Ensis goreensis* (Clessin, 1888) occurs in tropical West Africa, from the southern part of West Sahara to southern Angola (Lucira) and the Cape Verde Islands (Cosel 2009; and references therein). In Europe, there are four extant native *Ensis*, namely *E. magnus* Schumacher, 1817 [syn. *E. arcuatus* (Jeffreys, 1865)], *E. ensis* (Linné, 1758), *E. minor* (Chenu, 1843), and *E. siliqua* (Linné, 1758) (Cosel 2009). The American *E. directus* (Conrad, 1843) [syn. *E. americanus* (Gould, 1870)] is native to Atlantic North America, but was introduced to European coastal waters in 1978 (Cosel *et al.* 1982), where it now occurs from the North Sea to the Cantabrian Sea (Arias & Anadón 2012; Vierna *et al.* 2012). Another representative from Atlantic North America is the recently discovered *E. terranovensis* Vierna & Martínez-Lage, 2012, which was found off Newfoundland (Canada) and whose possible co-occurrence with *E. directus* is still unknown (Vierna *et al.* 2012). *Ensis macha* occurs along the southern coasts of Argentina, Peru, and Chile, and the remaining two taxa, *E. minor* Dall, 1899, and *E. minor megistus* Pilsbry & McGinty 1943, are native to the south-eastern coast of the USA (mainly Florida and the Gulf of Mexico).

The name *E. minor* Dall, 1899 is a junior homonym of the European *E. minor* (Chenu, 1843), but so far, without replacement name. Here, we propose the name *Ensis coseli* Vierna nom. nov. to replace *E. minor* Dall, 1899 non (Chenu, 1843). The new species name is in honour of Dr. Rudo von Cosel, for his contribution to the taxonomy and systematics of razor shells. *Ensis minor* Dall, 1899 has recently been renamed into *E. megistus* Pilsbry & McGinty, 1943, based on the synonymisation of *E. minor megistus* Pilsbry & McGinty, 1943 with the nominotypical subspecies (Huber 2010), which seems to make the proposition of a replacement name obsolete. However, after having studied the *E. minor megistus* type material and having compared it with specimens from some museum collections identified as *E. minor* sensu Dall (J. Vierna, A. M. González-Tizón, and A. Martínez-Lage, unpublished), we found a clear difference between both taxa in the position of the posterior adductor scar, which is a relevant taxonomic character in *Ensis* (see Cosel 2009; and Vierna *et al.* 2012). This scar is situated much more anterior to the pallial sinus in *E. minor megistus* than in the *E. minor* sensu Dall specimens examined. Therefore, we reject the synonymy in Huber (2010), which was provided without evidence, and we maintain *E. coseli coseli* and *E. coseli megistus* Pilsbry & McGinty, 1943 as subspecies, pending further studies, which may reveal species status for both. Molecular data of *E. coseli megistus* are currently unavailable, because to our best knowledge, there are no specimens available other than the valves that comprise the type material (see Pilsbry & McGinty 1943).

Despite their economic value, *Ensis* taxonomy is not well defined, due to the absence of clear autapomorphies in the shell and the sympatric distributions of some species. In fact, Cosel (2009) found some specimens that he considered as 'intergrades' of some European morphospecies, based on the analysis of the valves. This poorly defined taxonomy seems to be the cause of the mislabelling of specimens sold in European markets, where razor shells are often labelled as '*Ensis ensis*' even though they usually are either *E. magnus*, *E. siliqua*, *E. minor*, or *E. directus* (pers. obs).

We have noticed that several studies (Arias *et al.* 2011; Arias-Pérez *et al.* 2012; Varela *et al.* 2012; Rufino *et al.* 2012) have overlooked the fact that the European morphologically similar *E. minor* and *E. siliqua* occur sympatrically along Atlantic Europe. According to the review by Cosel (2009) based on shell morphology, *E. minor* occurs in both Atlantic and Mediterranean Europe, whereas *E. siliqua* is restricted to the Atlantic. However, the Galician Regional Government ('Xunta de Galicia', NW Spain), does not distinguish between these two taxa, even though fishing and commercialisation of these bivalves are strictly regulated in the area. Interestingly, González-Tizón *et al.* (2013) supported both taxa as separate species according to cytogenetics.

Phylogenetic analysis of DNA sequences is a common and useful approach for species delimitation (e.g. Fontaneto *et al.* 2011; Puillandre *et al.* 2012; Ornelas-Gatdula *et al.* 2012; Esselstyn *et al.* 2012) that should be considered in the light of other evidences such as morphology, biogeography, and ecology (integrative taxonomy, see Dayrat 2005). Even though the term 'species delimitation' is sometimes used in a similar way to 'DNA barcoding' (the use of standard genomic regions for identifying organisms to the species level, Hebert *et al.* 2003), for clarity reasons, in this paper, we make the distinction between these approaches.

Hebert *et al.* (2003) proposed to use a fragment of the so-called COI mitochondrial gene as the standard barcode for animals, but they recognised that supplemental analyses of one or more nuclear genes could be required when hybridisation or introgression occurs. Therefore, to delimit species and to select a suitable genomic region that can be used as a DNA barcode, in this paper, we analyse variation not only at COI, but also at other mitochondrial and nuclear regions, in more than one hundred specimens that were identified as belonging to the Atlantic morphospecies. If this work is taken into account by policymakers, it will contribute to the conservation and management of *Ensis* populations and will also enable a reliable labelling of the edible species, even in the absence of the valves (i.e. when processed as seafood), as well as a reliable identification of larvae and juvenile stages.

Materials and methods

Specimen identifications

We considered all currently known extant Atlantic *Ensis* morphospecies according to Cosel (2009) and Vierna *et al.* (2012). We studied a total of 109 specimens from 27 sampling sites (Table S1). Razor shell samples were obtained from museum collections and from colleagues, or they were directly sampled by us. Initial morphological identification of specimens was performed following Cosel (2009) and Vierna *et al.* (2012), and these were confirmed after phylogenetic analyses. Specimens were deposited in the collections of various natural history museums (see Table S1).

DNA isolation, PCR, and sequencing

Razor shell specimens were preserved either frozen or in 100% ethanol, except the *E. goreensis* sample. In this case, dry tissue was obtained from the interior part of the shell and rehydrated in sterile milli-Q water.

DNA was extracted from muscle tissue using the NucleoSpin Tissue kit (Macherey-Nagel GmbH and Co. KG, Düren, Germany). The mitochondrial regions studied were fragments of the cytochrome oxidase subunit I gene (COI) and the 16S ribosomal RNA gene (16S). The nuclear ribosomal DNA genes and spacers considered were the internal transcribed spacers 1 and 2 (ITS1 and ITS2), the 5.8S ribosomal gene (5.8S), and a fragment of the 18S ribosomal gene (18S). The ITS1-5.8S-ITS2 region was amplified, cloned, and sequenced as a whole. The COI region was sequenced in all 109 specimens, whereas the five other regions were sequenced in a subset of specimens (Table S1).

Using the COI 'universal' primers LCO1490 and HCO2198 (Folmer *et al.* 1994), we obtained some sequences that were then input in GeneFisher (Giegerich *et al.* 1996) to design four additional species-specific internal primer pairs (COI-Europe-1-F, 5′ GGG ATT AGT TGG GAC TAG G; COI-Europe-1-R, 5′ GTT AAA GCC CCT GCC AA; COI-Europe-2-F, 5′ TAG AGT TAG CTC GTC CT; COI-Europe-2-R, 5′ AAA TAG GGT CAC CAC CA; COI-*macha*-F, 5′ TAG TTG GGA CTA GGT TGA GA; COI-*macha*-R, 5′ TAG GAT CTC CTC CAC CTC T; COI-*coseli*-F, 5′ GAT TCG GTT AGA GTT AGC TCG A; and COI-*coseli*-R, 5′ GTT AAA GCA CCA GCT AGT ACA G). These new primers and the COI-*directus* ones (Vierna *et al.* 2012) were used in all COI amplification reactions. The primers used to amplify the 16S region were 16Sar and 16Sbr (Palumbi 1996). For the ITS1-5.8S-ITS2 region, we used the primers by Heath *et al.* (1995). Finally, for the 18S, we used the primers annealing at the 5′ and 3′ ends of the gene (Winnepenninckx *et al.* 1994). With the sequences obtained, we designed a pair of internal primers (18SintF, 5′ GAT CGT ACA ATC CTA CTT GG; and 18SintR, 5′ GCT CAT TAA CGG GAA CGA T) in GeneFisher. These new primers were employed in one of the *E. magnus* specimens analysed.

Each PCR (25 μL) contained ~25 ng of genomic DNA, 0.625 U of Taq DNA polymerase (Roche Diagnostics, Switzerland), 5 nmol of each dNTP (Roche Diagnostics), 20 pmol of each primer and the buffer recommended by the polymerase supplier. The general reaction conditions were as follows: an initial denaturation step at 94 °C for 3 min followed by 35 cycles of denaturation at 94 °C for 20 s; annealing at the following temperatures (LCO1490/HCO2198, 45 °C; COI-Europe-1, 57 °C; COI-Europe-2, 49 °C; COI-*macha*, 55 °C; COI-*coseli*, 52 °C; COI-*directus*, 48 °C; 16Sar/16Sbr, 44 °C; ITS1-5.8S-ITS2 region, 59 °C; 18S (Winnepenninckx *et al.* 1994), 56 °C; and 18Sint, 55 °C) for 20 s; extension at 72 °C for 30–50 s; and a final extension at 72 °C for 5 min. PCR products were run on 1 % agarose gels, stained with either ethidium bromide or Real Safe (Real, Valencia, Spain) and imaged under UV light.

All PCRs yielded single-band patterns, and therefore, amplicons were directly sequenced (except the ITS1-5.8S-ITS2 region ones), after being purified with ExoSAP-IT (USB, Santa Clara, CA, USA). PCR primers were used to sequence amplicons in both directions. In the case of 18S amplicons, we designed internal sequencing primers (18SseqF, 5′ CCC GTA ATT GGA ATG AGT AC; and 18SseqR, 5′ CGA ATC AAG AAA GAG CTC TC) to cover the whole amplified region. Due to the occurrence of intragenomic variation (Vierna *et al.* 2010), ITS1-5.8S-ITS2 PCR products were cloned using the TOPO TA cloning kit (Invitrogen, Paisley, UK). Transformant colonies were selected, and insert size was checked by PCR.

We spread one clone per individual on an LB plate and let it grow overnight at 37 °C. Plasmids were purified with the QiaPrep Spin Miniprep Kit (Qiagen, Hilden, Germany), and they were sequenced using the M13 forward and reverse primers (supplied with the cloning kit). In addition to the sequences generated, some *E. directus* and *E. terranovensis* COI, ITS1, and ITS2 sequences (Vierna *et al.* 2012) from specimens sampled at Cobscook Bay, Long Pond, and The Wash were included in our data sets (see Table S2 for accession numbers). New DDBJ/EMBL/GenBank accession numbers are HF970346-HF970575 and HF975604-HF975627.

Bioinformatic analyses

The software BioEdit 7.0.9.0 (Hall 1999) and Geneious Pro 5.4.6 (Drummond *et al.* 2011) were used to examine the electropherograms. To search for stop codons that would be indicative of the presence of pseudogenes, the COI amino acid sequences were obtained from MEGA 5.03 (Tamura *et al.* 2011) using the 'invertebrate mitochondrial genetic code'.

COI and 16S alignments were carried out in ClustalW 2.0 (Larkin *et al.* 2007) under default parameters. ITS1 and ITS2 sequences were aligned using the Q-INS-i strategy as implemented in MAFFT, version 7 (Katoh & Toh 2008; Katoh & Standley 2013). Highly variable regions were deleted from ITS1 and ITS2 alignments using Gblocks (Castresana 2000; Talavera & Castresana 2007) available at http://molevol.cmima.csic.es/castresana/Gblocks_server.html under default (conservative) options. All alignments were trimmed, and identical sequences were collapsed into sequence-types using DnaSP 5.10.01 (Librado & Rozas 2009). Multigene alignments (COI+16S; ITS1+ITS2) were obtained as follows: we aligned each genomic region independently; then, we concatenated the sequences obtained from each specimen, and finally, we collapsed identical multigene sequences into sequence-types.

Phylogenetic trees were inferred under Bayesian (BA) and maximum-likelihood (ML) methods. BA was carried out using the software MrBayes 3.1.2 (Huelsenbeck & Ronquist 2001) from the CIPRES Science Gateway (Miller *et al.* 2010). Models of evolution for each partition (Table S3) were obtained from MrModelTest 2.3 (Johan Nylander, http://www.abc.se/~nylander/). The analysis was performed with 15 000 000 generations initiated with a random starting tree, sampling every 1000 generations, and allowing the program to estimate the likelihood parameters required. Stationarity was assessed using the Web-based software AWTY (Nylander *et al.* 2008). Results collected prior to stationarity were discarded as burn-in. ML phylogenies were obtained using RAxML 7.2.8 (Stamatakis 2006; Stamatakis *et al.* 2008) that was run from the same bioinformatic platform. This software is capable of assigning and estimating separate model parameters for individual genes of multigene alignments (Stamatakis 2006) and implements the general time-reversible (GTR) substitution model for all partitions. Node confidence was assessed using 1000 nonparametric bootstrap replicates (Felsenstein 1985). All phylogenetic trees were edited in Dendroscope 3 (Huson & Scornavacca 2012).

The best resolved phylogenetic trees based on the CO1+16S and ITS1+ITS2 data sets were subjected to a general mixed Yule coalescent analysis (GMYC) (Pons *et al.* 2006; Fontaneto *et al.* 2007). The model was performed with R 2.15.3 (R Development Core Team 2011) package splits (SPecies' LImits by Threshold Statistics) available at http://r-forge.r-project.org/projects/splits/. The required ultrametric phylogenetic trees were generated using penalised likelihood in r8s 1.70 and mid-point rooted (Sanderson 2003; Obertegger *et al.* 2012).

Even though we considered the possibility of including *Pharus legumen* (Linné, 1758) and *Siliqua patula* (Dixon, 1789) sequences as outgroups, they were too much divergent to reliably align them in the case of some genomic regions (ITS1 and ITS2). Therefore and considering that it is not necessary to use outgroups to check whether morphospecies are separately evolving lineages (e.g. Ley & Hardy 2010; James *et al.* 2010), we decided against using outgroup sequences, and thus, all the phylogenetic trees we present here are unrooted.

We also obtained phylogenetic networks for each alignment using the Neighbor-Net algorithm (Bryant & Moulton 2004) and uncorrected p-distances in SplitsTree4 (Huson & Bryant 2006). Finally, the mean intraspecific and interspecific uncorrected p-distances were obtained from MEGA 5.03. Their standard errors were calculated by 1000 bootstraps.

Results

In general, morphological identifications were confirmed by molecular data. However, the morphologically similar *E. minor* and *E. siliqua* were rather difficult to identify reliably in terms of shell morphology using available keys (Cosel 2009). Nonetheless, we were able to clearly differentiate them from the other European morphospecies *E. magnus* and *E. ensis*. Therefore, they were identified both using phylogenetic clustering and geographic information: because according to Cosel (2009), *E. siliqua* is restricted to the Atlantic, whereas *E. minor* occurs both in the Atlantic and in the Mediterranean, the Atlantic specimens that clustered with the Mediterranean ones were considered as *E. minor*.

Interestingly, two specimens from Jacksonville (FL, USA) that we received from the Florida Museum of

Natural History, labelled as *E. minor* Dall, 1899 (*E. coseli coseli*), resulted as belonging to *E. directus*, according to the shape of muscle scars on the inner valves (Cosel 2009) and to phylogenetic clustering. These specimens are, to our knowledge, the first confirmed record of *E. directus* as far south as north-eastern Florida.

The COI alignment (303 bp) did not display, as expected, any gap. No stop codons were found, and there were only three mutations (from three different sequences) that lead to amino acid substitutions. The proportion of variable sites (s) for this region was $s = 0.337$. Alignments of 16S (340 bp), 18S (1271 bp) and 5.8S (157 bp) sequences were straightforward. The proportion of variable sites was $s = 0.126$, $s = 0.009$ and $s = 0.032$, respectively. There was only one gap position, in the 18S region.

Regarding the alignments of ITS1 and ITS2 sequences, 285 (44 % of the original 636 positions) were retained after the Gblocks filtering in the case of ITS1. For ITS2, 228 (62 % of the original 364 positions) were kept after filtering. The proportion of polymorphic sites after filtering was $s = 0.154$ and $s = 0.224$, respectively.

The number of haplotypes obtained from the COI alignment was 81 of 109 sequences (74.3 %), whereas the 16S alignment yielded 29 haplotypes of 80 sequences (36.2 %). The 18S region produced six sequence-types of 13 sequences (46.1 %), the 5.8S region yielded only five sequence-types of 41 sequences (12.2 %), and both ITS1 and ITS2 regions yielded 19 of 41 sequences (46.3 %) each. The COI+16S multigene alignment produced 70 haplotypes of 80 sequences (87.5 %), whereas the ITS1+ITS2 alignment yielded 29 sequence-types of 41 sequences (70.7 %).

Sequence-type distributions of the 18S and 5.8S regions (that due to their conservation, lacked resolution at the species level and were not considered in the subsequent analyses) were as follows: *E. magnus*, *E. siliqua*, and *E. minor* shared the same 18S sequence-type. However, *E. directus*, *E. terranovensis*, *E. macha*, *E. coseli coseli*, and *E. ensis* had one (non-shared) sequence-type each. The two individuals of *E. macha* from Chile and Argentina whose 18S region was sequenced displayed the same sequence. In the case of the 5.8S region, one of the five sequence-types obtained was shared by all North and South American species. Another one was shared by all European species. The remaining three corresponded to one *E. terranovensis*, one *E. directus*, and one *E. magnus* sequence.

All phylogenetic trees obtained, based either on mitochondrial or on nuclear DNA, recovered morphospecies at each side of the Atlantic as reciprocally monophyletic (Figs 1 and 2; Figs S1-S6).

Trees based on the COI alignment recovered all morphospecies as monophyletic, with the exception of *E. terranovensis* in the BA analysis, and both *E. terranovensis* and *E. magnus* in the ML analysis (Fig. S1). The morphospecies *E. goreensis* was represented by only one haplotype.

The 16S alignment produced one haplotype that was shared between *E. terranovensis* and *E. directus*. Both BA and ML phylogenetic trees clustered *E. magnus* and *E. ensis* sequences on monophyletic groups. As for *E. coseli coseli* and *E. goreensis*, only one haplotype was obtained per morphospecies. In the BA analysis, sequences from *E. macha*, *E. siliqua*, and *E. minor* were clustered in monophyletic groups as well (see Fig. S2).

The BA tree reconstructed from the COI+16S multigene alignment recovered all morphospecies as monophyletic, except for one *E. terranovensis* haplotype, which formed a polytomy (Fig. 1). The support values for each of those clades ranged between 0.98 and 1.00 except for *E. terranovensis* and *E. directus*. The ML analysis recovered as monophyletic all the European and African species with the exception of *E. ensis*. In contrast, the sequences from American species did not cluster, in general, into monophyletic groups in the ML tree. Rather, they formed some polytomies, with the exception of the *E. coseli coseli* sequence-types (see Fig. S3).

No *E. goreensis* ITS1-5.8S-ITS2 sequences could be obtained, probably due to degradation of the extracted DNA, and therefore, we lack nuclear DNA data for this morphospecies. Regarding the ITS1 analysis, sequences from *E. macha*, *E. siliqua*, *E. magnus*, and *E. minor* yielded only one sequence-type per species. Sequences from the other morphospecies were clustered into monophyletic groups, in both BA and ML phylogenetic trees (see Fig. S4).

In the ITS2 analysis, *E. siliqua*, *E. magnus*, and *E. ensis* yielded one sequence-type each. The BA tree recovered all morphospecies as monophyletic with the exception of *E. directus* and *E. terranovensis*, which formed two polytomies. These two species showed a paraphyletic pattern in the ML phylogeny, whereas all others were recovered as monophyletic (see Fig. S5).

Finally, in the ITS1+ITS2 analysis, there were two morphospecies (*E. magnus* and *E. siliqua*) that yielded only one sequence-type each. The BA phylogenetic tree recovered all other morphospecies as monophyletic, with the exception of *E. terranovensis*, which formed a polytomy. Support values for each of these clades ranged between 0.95 and 1.00. In the ML tree, all morphospecies were recovered as monophyletic, and support values were high (94–100) except for *E. terranovensis* and *E. minor* (see Fig. 2; Fig. S6).

The GMYC analysis based on the BA COI+16S phylogenetic tree revealed a model with eight different entities (confidence interval 8–11) as the most likely (likelihood of the null model, 338.693; maximum likelihood of the

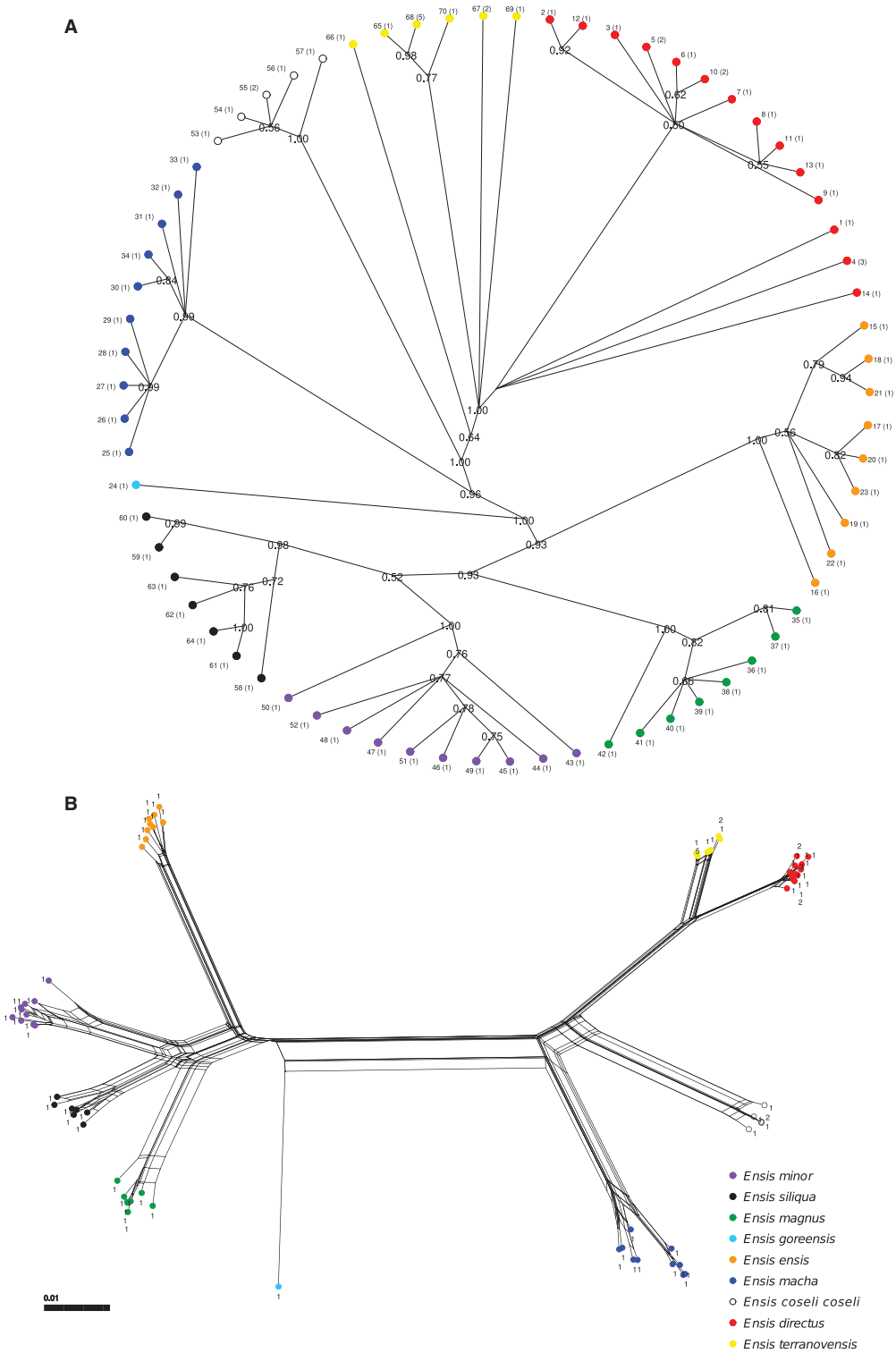


Fig. 1 Phylogenetic relationships among mitochondrial (COI+16S) haplotypes. Haplotype frequencies are shown in parentheses in the tree, and at each terminal node, in the network. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from. —A. Unrooted Bayesian phylogenetic tree. Node confidence values below 0.5 are not shown. —B. Phylogenetic network constructed using the Neighbor-Net algorithm and uncorrected p-distances.

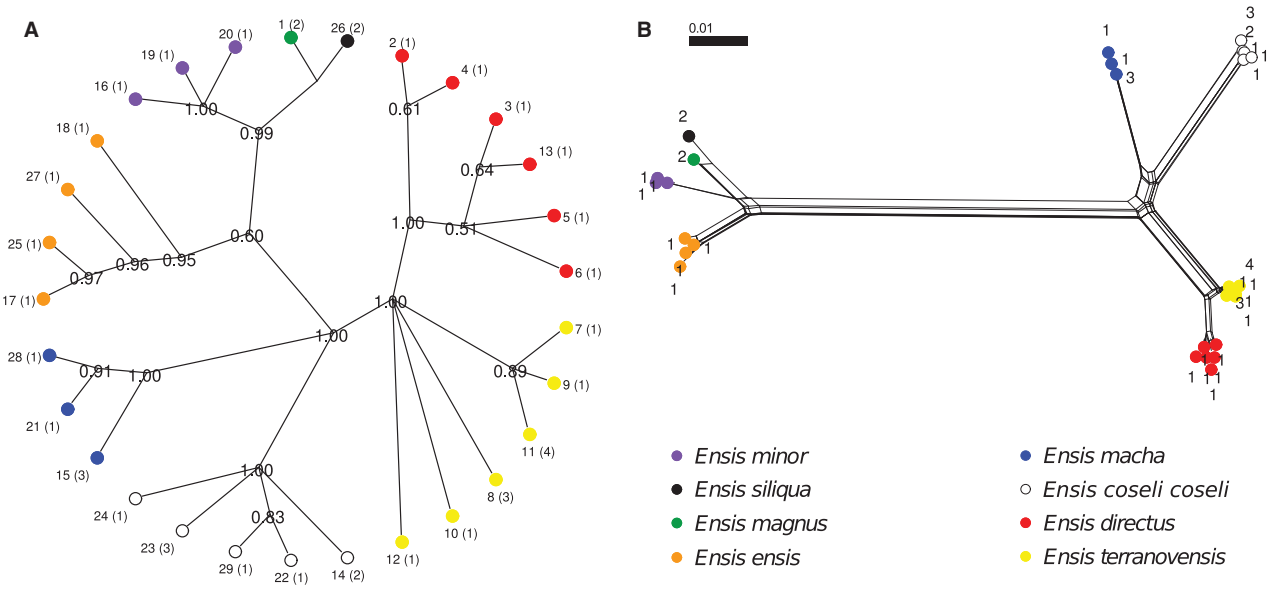


Fig. 2 Phylogenetic relationships among nuclear (ITS1+ITS2) sequence-types. Sequence-type frequencies are shown in parentheses in the tree, and at each terminal node, in the network. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from. —A. Unrooted Bayesian phylogenetic tree. Node confidence values below 0.5 are not shown. —B. Phylogenetic network constructed using the Neighbor-Net algorithm and uncorrected p-distances.

GMYC model, 352.236; likelihood ratio, 27.085; LR test < 0.0001). The number of morphospecies included in the analysis (nine) falls within the confidence interval. Therefore, the GMYC analysis based on the COI+16S phylogeny is consistent with the existence of the nine different species.

The GMYC analysis based on the ML ITS1+ITS2 phylogenetic tree revealed a model with eight different entities (confidence interval 7–8) as the most likely (likelihood of the null model, 113.846; maximum likelihood of the GMYC model, 120.472; likelihood ratio, 13.253; result of the LR test, 0.004). In the same way as for the CO1+16S phylogeny, the GMYC analysis based on the ITS1+ITS2 data set is consistent with the existence of eight different species (in this case, there were no *E. goreensis* sequences available).

The phylogenetic networks obtained based on the COI, COI+16S, ITS1, and ITS1+ITS2 alignments clearly supported the existence of groups that perfectly matched the currently described morphospecies (Figs S1 and S4; Figs 1 and 2).

The 16S region was not able to completely differentiate among North American species because, as already mentioned, there was one haplotype shared by two of them. Regarding European species, the two *E. siliqua* haplotypes were not clustered together in the network (see Fig. S2).

The ITS2 phylogenetic network clustered the sequence-types obtained from *E. terranovensis* and *E. directus* specimens and also the ones obtained from *E. siliqua*, *E. magnus*, and *E. ensis* (see Fig. S5).

We calculated the intraspecific and interspecific mean p-distance values for the COI, 16S, ITS1, and ITS2 regions using the trimmed alignments containing all sequences (i.e. not the sequence-type alignments; Tables S4–S7).

We also calculated interpopulation mean p-distances for *E. macha* individuals from Puerto Lobos (Argentina) and Playa Dichato (Chile). In this case, no ITS1 or ITS2 values were obtained because the number of sequences available was too small. Divergence between the two *E. macha* populations was in some cases more pronounced than interspecific divergence. For instance, 16S interpopulation divergence was 0.011 ± 0.004, whereas interspecific divergence in some species pairs was equal or lower: *E. directus* and *E. terranovensis*, 0.003 ± 0.003; and *E. siliqua* and *E. magnus*, 0.011 ± 0.005 (Table S5). Similarly, COI interpopulation divergence was 0.038 ± 0.010, quite comparable to that calculated for *E. directus* and *E. terranovensis* (0.058 ± 0.012) and for *E. siliqua* and *E. magnus* (0.064 ± 0.012; Table S4).

Discussion

Species delimitation and molecular systematics

Our results support that the nine morphospecies under study are different evolutionary lineages, according to both nuclear and mitochondrial DNA, and in agreement with Cosel (2009). Therefore, they should be considered as nine different species.

The split between *E. directus* and *E. terranovensis* was already demonstrated by Vierna *et al.* (2012) using

morphometrics and nucleotide variation at COI, ITS1, ITS2, and ANT (a fragment of a nuclear single-copy region coding adenine nucleotide translocase). However, in the present work, the split between the two species was not evident for all genomic regions analysed, because in some phylogenetic trees *E. terranovensis*, sequences formed polytomies instead of clustering in a clade. These polytomies could reflect that the two sister species underwent speciation rather recently, a fact that is further supported by the shared 16S haplotype, which is an example of incomplete lineage sorting among closely related taxa.

Despite the absence of tropical west American *Ensis* in our analyses, we have unveiled a strong phylogeographic structure within genus *Ensis*. Species at each side of the Atlantic are reciprocally monophyletic, as shown by molecular data as well as for the shape of the pallial sinus (broad and shaped like an irregular W in American species and narrower and more or less rounded in European species, Cosel 2009).

In Atlantic North America, three lineages were found: whereas *E. directus* and *E. terranovensis* were supported as sister species, the position of *E. coseli coseli* was variable. This taxon branched off as sister to *E. directus* and *E. terranovensis* in most of the analyses, supporting the monophyly of North American species, and as sister to *E. macha*, in some others. Interestingly, we have confirmed the occurrence of *E. directus* as south as north-eastern Florida.

In South America, *E. macha* specimens sampled at Playa Dichato and Puerto Lobos were rather divergent, again showing phylogeographic structure. The number of individuals analysed was small and prevents any general conclusion about possible population structure of these razor shells along their distributional ranges. However, the degree of divergence displayed by individuals from Playa Dichato and Puerto Lobos according to COI and 16S was in some cases even higher than that obtained for some other *Ensis* interspecific comparisons, and suggests incipient speciation. Significant population structure has also been found in marine gastropod *Nacella magellanica* (Gmelin, 1791) populations from Atlantic and Pacific Patagonia (González-Wevar *et al.* 2012). Taken into account that Camus (2001) identified a major biogeographic break in Chile at 41–43°S and that Playa Dichato is located in the north of these parallels, this biogeographic break could also explain population structure in *E. macha*. Our results highlight the need to study the population structure and diversity of this razor shell along its entire distribution (a survey on shell morphometry along the Patagonian coast is available, Márquez & Van der Molen 2011) to design conservation plans for this intensively fished species (see Hernández *et al.* 2011).

Unfortunately, the phylogenetic relationships between species within the European/African clade were not well resolved as evidenced by the presence of polytomies and low node support values.

Remarkably, we have demonstrated that *E. siliqua* and *E. minor* are different species, despite their morphological similarities. Our results agree with the taxonomic revision by Cosel (2009) based on shell morphology and with González-Tizón *et al.* (2013) who found important differences among the karyotypes of both species.

Some of the specimens studied in this work were also analysed in two previous papers on evolutionary genetics. According to our results, some sequences that were thought to have been obtained from *E. siliqua* specimens (Vierna *et al.* 2009, 2011) were in fact obtained from *E. minor* razor shells. The sequence labels have been updated on DDBJ/EMBL/GenBank. Fortunately, we do not expect that this misidentification of specimens impacts any of the main conclusions of those papers. However, this issue accentuates the importance of DNA barcoding in research fields other than ecology and biodiversity.

Given that both *E. siliqua* and *E. minor* co-occur in Atlantic Europe and that it is not always possible to reliably differentiate among specimens based on a visual inspection of the shell, we suspect that in some previous reports on *E. siliqua* (Darriba *et al.* 2005; Arias *et al.* 2011; Arias-Pérez *et al.* 2012; Varela *et al.* 2012; Rufino *et al.* 2012), specimens from both species could have been confused. The authors presumably determined specimens based on shell morphology, although details on how identifications were carried out were not provided.

Even though Varela *et al.* (2012) found differentiation among northern and southern '*E. siliqua*' populations, Arias *et al.* (2011) and Arias-Pérez *et al.* (2012) found a Portuguese population clustering with an Irish population. In the morphometric study by Rufino *et al.* (2012), they described three different 'morphs' based on shell morphometrics, one from an Irish population and two others from Portuguese populations. Taking our results into account, the differentiation that they found among populations might have been produced by a two-species scenario, rather than by population isolation. This can be tested by DNA barcoding some of the specimens analysed in those surveys.

Due to sampling limitations, it was out of the scope of this work to investigate the distribution ranges of species in detail. Nonetheless, because we found *E. siliqua* in Borkum reef, Central North Sea, and Cedeira, and *E. minor* in Bandol, La Capte, Lira, and Ría de Vigo, one might think that *E. minor* is distributed from the central Galician 'rías' to the Mediterranean, whereas *E. siliqua* has a more northern distribution. However, the distribution range of *E. minor* north of Galicia and northward to Scotland was confirmed by Cosel (2009) based on shell morphology of his own samplings, as were different biotopes and the sympatric

occurrence of *E. siliqua* in the same areas. Only in the southern part of Brittany Peninsula, were a few morphological intergrades of *E. minor* x *E. siliqua* found (Cosel 2009).

DNA barcodes for Atlantic Ensis

The ideal DNA barcode should meet the following conditions: (1) suitable primer pairs should be available, (2) sequence-types should not be shared even by specimens from closely related species, (3) alignments should be straightforward, (4) intragenomic variation should be low to avoid cloning of PCR products before sequencing, and (5) sequences obtained from individuals from the same species should cluster together (and apart from sequences obtained from other species) in a phylogenetic tree or network.

In Atlantic *Ensis*, condition (1) was met by all genomic regions considered, even though in the case of COI, several primer pairs were used. Condition (2) was not met by 5.8S nor by 18S regions so they were not considered in subsequent phylogenetic analyses. Even though 5.8S is not widely used in phylogenetics due to its short length and scarce polymorphism, the 18S region is a popular phylogenetic marker. In *Ensis*, 18S was useful in differentiating the European/African clade from the American clade, but this region failed to differentiate among closely related species, in agreement with Tang *et al.* (2012). Despite the fact that the 16S region did not meet condition (2), it was a useful marker for species delimitation. Conditions (3) and (4) were not met by ITS1 nor by ITS2, as already reported by Vierna *et al.* (2010), but were suitable for species delimitation. Finally, even though condition (5) depends on the particular phylogenetic analyses carried out, it was met by COI and ITS1 in most cases (very clearly in phylogenetic networks, but not as clear in some phylogenetic trees, especially for *E. directus* and *E. terranovensis* sequences). Therefore, we suggest using COI as the main DNA barcode in Atlantic *Ensis* species.

Even though some authors have attempted to establish molecular identification protocols for some *Ensis* using PCR and PCR-RFLPs (Fernández-Tajes & Méndez 2007; Freire *et al.* 2008; Fernández-Tajes *et al.* 2010), because species boundaries were still unclear at that time, the proposed methodologies might give misleading results in some cases. Moreover, DNA barcoding is a much more informative (and therefore reliable) approach compared with PCR/PCR-RFLP based methods, even though the cost is somewhat higher due to the additional sequencing step.

Conclusion and future directions

In this work, using mitochondrial and nuclear DNA sequence data, we have clarified the species limits within Atlantic *Ensis*. In addition, we show that DNA barcoding is capable of identifying *Ensis* specimens to the species level and suggest taking advantage of this methodology for reliable and inexpensive identifications of specimens.

Finally, we encourage regional authorities to study the species composition of razor shell communities and establish specific conservation and managing plans, at least in the areas where *Ensis* spp. are an economic resource.

Acknowledgements

We thank very much the following colleagues and institutions that helped us in some way or another: A. Anker and co-workers, Ana de la Torriente, Alejandro Martínez, Anja Schulze, Antón Vizcaíno, Benthic Team at IMARES Wageningen UR, Diana Fernández, Diego Fontaneto, David Palmer, Emilie Egea, José María Casariego, José Fernández, Katrine Worsaae, Iben Heiner, Javier De Andrés, John Slapcinsky, Jon Havenhand, L. Kirkendale, Lobo Orensanz, M. Pin, Meromar Seafoods BV, Nicolas Puillandre, Philip Sargent, Pablo Pita, Rafael Araujo, Ray J. Thompson, R. O'Donnell, Rüdiger Schmelz, Rudo von Cosel, Stephen Tettelbach, Tim Sheehan (Gulf of Maine).

References

Arias, A. & Anadón, N. (2012). First record of *Mercenaria mercenaria* (Bivalvia: Veneridae) and *Ensis directus* (Bivalvia: Pharidae) on Bay of Biscay, Iberian Peninsula. *Journal of Shellfish Research*, *31*, 57–60.

Arias, A., Fernández-Moreno, M., Fernández-Tajes, J., Gaspar, M. B. & Méndez, J. (2011). Strong genetic differentiation among east Atlantic populations of the sword razor shell (*Ensis siliqua*) assessed with mtDNA and RAPD markers. *Helgoland Marine Research*, *65*, 81–89.

Arias-Pérez, A., Fernández-Tajes, J., Gaspar, M. B. & Méndez, J. (2012). Isolation of microsatellite markers and analysis of genetic diversity among east Atlantic populations of the sword razor shell *Ensis siliqua*: a tool for population management. *Biochemical Genetics*, *50*, 397–415.

Ariz Abarca, L., Cortés Segovia, C., González Yáñez, J., Barahona Toledo, N. & Nilo Gatica, M. (2007). Situación actual de la pesquería del recurso huepo (*Ensis macha*) en la VII Región. Instituto de Fomento Pesquero (Accessed at http://www.fip.cl/FIP/Archivos/pdf/informes/inffinal%202006-44.pdf).

Barón, P. J., Real, L. E., Ciocco, N. F. & Ré, M. E. (2004). Morphometry, growth and reproduction of an Atlantic population of the razor clam *Ensis macha* (Molina, 1782). *Scientia Marina*, *68*, 211–217.

Bryant, D. & Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, *21*, 255–265.

Camus, P. A. (2001). Biogeografía marina de Chile continental. *Revista Chilena de Historia Natural*, *74*, 587–617.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*, 540–552.

Cosel, von R. (2009). The razor shells of the eastern Atlantic, part 2. Pharidae II: the genus *Ensis* Schumacher, 1817 (Bivalvia, Solenoidea). *Basteria*, *73*, 1–48.

Cosel, von R., Dörjes, J. & Mühlenhardt-Siegel, U. (1982). Die amerikanische Schwertmuschel *Ensis directus* (Conrad) in der Deutschen Bucht. I. Zoogegraphis und Taxonomie im Vergleich mit den einheimischen Schwertmuschel-Arten. *Senckenbergiana maritima*, *14*, 147–173.

Darriba, S., San Juan, F. & Guerra, A. (2005). Gametogenic cycle of *Ensis siliqua* (Linnaeus, 1758) in the Ría de Corcubión, northwestern Spain. *Journal of Molluscan Studies*, *71*, 47–51.

Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society*, *85*, 407–415.

Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T. & Wilson, A. (2011). Geneious v5.4. Available from http://www.geneious.com/.

Esselstyn, J. A., Evans, B. J., Sedlock, J. L., Khan, F. A. A. & Heaney, L. R. (2012). Single-locus species delimitation: a test of the mixed Yule – coalescent model, with an empirical application to Philippine round-leaf bats. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 3678–3686.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, *39*, 783–791.

Fernández-Tajes, J. & Méndez, J. (2007). Identification of the razor clam species *Ensis arcuatus*, *E. siliqua*, *E. directus*, *E. macha*, and *Solen marginatus* using PCR-RFLP analysis of the 5S rDNA region. *Journal of Agricultural and Food Chemistry*, *55*, 7278–7282.

Fernández-Tajes, J., Freire, R. & Méndez, J. (2010). A simple one-step PCR method for the identification between European and American razor clams species. *Food Chemistry*, *118*, 995–998.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*, 294–299.

Fontaneto, D., Herniou, E. A., Boschetti, C., Caprioli, M., Melone, G., Ricci, C. & Barraclough, T. G. (2007). Independently evolving species in asexual bdelloid rotifers. *PLoS Biology*, *5*, e87.

Fontaneto, D., Iakovenko, N., Eyres, I., Kaya, M., Wyman, M. & Barraclough, T. G. (2011). Cryptic diversity in the genus Adineta Hudson & Gosse, 1886 (Rotifera: Bdelloidea: Adinetidae): a DNA taxonomy approach. *Hydrobiologia*, *662*, 27–33.

Freire, R., Fernández-Tajes, J. & Méndez, J. (2008). Identification of razor clams *Ensis arcuatus* and *Ensis siliqua* by PCR–RFLP analysis of ITS1 region. *Fisheries Science*, *74*, 511–515.

Giegerich, R., Meyer, F. & Schleiermacher, C. (1996). GeneFisher - software support for the detection of postulated genes. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, *4*, 68–77.

González-Tizón, A. M., Rojo, V., Vierna, J., Jensen, K. T., Egea, E. & Martínez-Lage, A. (2013). Cytogenetic characterisation of the razor shells *Ensis directus* (Conrad, 1843) and *E. minor* (Chenu, 1843) (Mollusca: Bivalvia). *Helgoland Marine Research*, *67*, 73–82.

González-Wevar, C. A., Hüne, M., Cañete, J. I., Mansilla, A., Nakano, T. & Poulin, E. (2012). Towards a model of postglacial biogeography in shallow marine species along the Patagonian Province: lessons from the limpet *Nacella magellanica* (Gmelin, 1791). *BMC Evolutionary Biology*, *12*, 139.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Heath, D. D., Rawson, P. D. & Hilbish, T. J. (1995). PCR-based nuclear markers identify alien blue mussel (*Mytilus* spp.) genotypes on the west coast of Canada. *Canadian Journal of Fisheries and Aquatic Sciences*, *52*, 2621–2627.

Hebert, P. D., Cywinska, A. & Ball, S. L. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, 313–321.

Hernández, A. F., Cubillos, L. A. & Quiñones, R. A. (2011). Evaluación talla estructurada de los stocks de *Ensis macha* y *Tagelus dombeii* en el Golfo de Arauco, Chile. *Revista de Biología Marina y Oceanografía*, *46*, 157–176.

Huber, M. (2010). *Compendium of Bivalves. A Full-Color Guide to 3,300 of the World's Marine Bivalves. A Status on Bivalvia after 250 Years of Research*. Hackenheim, Germany: ConchBooks.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*, 754–755.

Huson, D. H. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*, 254–267.

Huson, D. H. & Scornavacca, C. (2012). Dendroscope 3: an interactive viewer for rooted phylogenetic trees and networks. *Systematic Biology*, *61*, 1061–1067.

James, S. W., Porco, D., Decaëns, T., Richard, B., Rougerie, R. & Erséus, C. (2010). DNA barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): resurrection of *L. herculeus* (Savigny, 1826). *PLoS One*, *5*, e15629.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780.

Katoh, K. & Toh, H. (2008). Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*, *9*, 212.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., López, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*, 2947–2948.

Ley, A. C. & Hardy, O. J. (2010). Species delimitation in the Central African herbs *Haumania* (Marantaceae) using georeferenced nuclear and chloroplastic DNA sequences. *Molecular Phylogenetics and Evolution*, *57*, 859–867.

Librado, P. & Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*, 1451–1452.

Márquez, F. & Van der Molen, S. (2011). Intraspecific shell-shape variation in the razor clam *Ensis macha* along the Patagonian coast. *Journal of Molluscan Studies*, *77*, 123–128.

Miller, M. A., Pfeiffer, W. & Schwartz, T. (2010), 'Creating the CIPRES Science Gateway for inference of large phylogenetic trees' In *Proceedings of the Gateway Computing Environments Workshop (GCE)* (pp. 1–8). 14 Nov. 2010, New Orleans, LA.

Nylander, J. A. A., Wilgenbusch, J. C., Warren, D. L. & Swofford, D. L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, *24*, 581–583.

Obertegger, U., Fontaneto, D. & Flaim, G. (2012). Using DNA taxonomy to investigate the ecological determinants of plankton diversity: explaining the occurrence of *Synchaeta* spp. (Rotifera, Monogononta) in mountain lakes. *Freshwater Biology*, *57*, 1545–1553.

Ornelas-Gatdula, E., Camacho-García, Y., Schrödl, M., Padula, V., Hooker, Y., Gosliner, T. M. & Valdés, A. (2012). Molecular systematics of the '*Navanax aenigmaticus*' species complex (Mollusca, Cephalaspidea): coming full circle. *Zoologica Scripta*, *41*, 374–385.

Palumbi, S. R. (1996). Nucleic acids II: the polymerase chain reaction. In D. M. Hillis, C. Moritz & B. K. Mable (Eds) *Molecular Systematics* (pp. 205–247). Sunderland, MA: Sinauer & Associates Inc.

Pilsbry, H. A. & McGinty, T. L. (1943). *Ensis minor megistus* n. subsp., a West Florida razor clam. *The Nautilus*, *57*, 33–34.

Pons, J., Barraclough, T. G., Gómez-Zurita, J., Cardoso, A., Durán, D. P., Hazell, S., Kamoun, S., Sumlin, W. D. & Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, *55*, 595–609.

Puillandre, N., Modica, M. V., Zhang, Y., Sirovich, L., Boisselier, M. C., Cruaud, C., Holford, M. & Samadi, S. (2012). Large-scale species delimitation method for hyperdiverse groups. *Molecular Ecology*, *21*, 2671–2691.

R Development Core Team (2011) *R Foundation for Statistical Computing*, Vol. 1, issue: 2.11.1.

Rufino, M. M., Vasconcelos, P., Pereira, F., Fernández-Tajes, J., Darriba, S., Méndez, J. & Gaspar, M. B. (2012). Geographical variation in shell shape of the pod razor shell *Ensis siliqua* (Bivalvia: Pharidae). *Helgoland Marine Research*, *67*, 49–58. DOI 10.1007/s10152-012-0303-6.

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, *19*, 301–302.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*, 2688–2690.

Stamatakis, A., Hoover, P. & Rougemont, J. (2008). A fast bootstrapping algorithm for the RAxML web-servers. *Systematic Biology*, *57*, 758–771.

Talavera, G. & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*, 564–577.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*, 2731–2739.

Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T. G. & Fontaneto, D. (2012). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences*, *109*, 16208–16212.

Varela, M. A., Martínez-Lage, A. & González-Tizón, A. M. (2012). Genetic heterogeneity in natural beds of the razor clam *Ensis siliqua* revealed by microsatellites. *Journal of the Marine Biological Association of the United Kingdom*, *92*, 171–177.

Vierna, J., González-Tizón, A. M. & Martínez-Lage, A. (2009). Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). *Biochemical Genetics*, *47*, 635–644.

Vierna, J., Martínez-Lage, A. & González-Tizón, A. M. (2010). Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. *Genome*, *53*, 23–34.

Vierna, J., Jensen, K. T., Martínez-Lage, A. & González-Tizón, A. M. (2011). The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae). *Heredity*, *107*, 127–142.

Vierna, J., Jensen, K. T., González-Tizón, A. M. & Martínez--Lage, A. (2012). Population genetic analysis of *Ensis directus* unveils high genetic variation in the introduced range and reveals a new species from the NW Atlantic. *Marine Biology*, *159*, 2209–2227.

Winnepenninckx, B., Backeljau, T. & De Wachter, R. (1994). Small ribosomal subunit RNA and the phylogeny of Mollusca. *The Nautilus Supplement*, *2*, 98–110.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Phylogenetic relationships among COI haplotypes.

Fig. S2. Phylogenetic relationships among 16S haplotypes.

Fig. S3. Unrooted maximum likelihood phylogenetic tree based on the COI+16S haplotype alignment.

Fig. S4. Phylogenetic relationships among ITS1 sequence-types.

Fig. S5. Phylogenetic relationships among ITS2 sequence-types.

Fig. S6. Unrooted maximum likelihood phylogenetic tree based on the ITS1+ITS2 sequence-type alignment.

Table S1. Razor shell specimens studied.

Table S2. Accession number of sequences retrieved from Vierna *et al.* (2012).

Table S3. Models of evolution for each partition.

Table S4. Mean interspecific and mean intraspecific divergence based on the COI alignment.

Table S5. Mean interspecific and mean intraspecific divergence based on the 16S alignment.

Table S6. Mean interspecific and mean intraspecific divergence based on the ITS1 alignment.

Table S7. Mean interspecific and mean intraspecific divergence based on the ITS2 alignment.

Fig. S1 Phylogenetic relationships among COI haplotypes. Haplotype frequencies are shown in parentheses in the tree, and at each terminal node, in the network. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from. **A** Unrooted Bayesian phylogenetic tree. Node confidence values below 0.5 are not shown. **B** Unrooted maximum likelihood phylogenetic tree. Node confidence values below 50 are not shown. **C** Phylogenetic network constructed using the neighbournet algorithm and uncorrected p-distances.

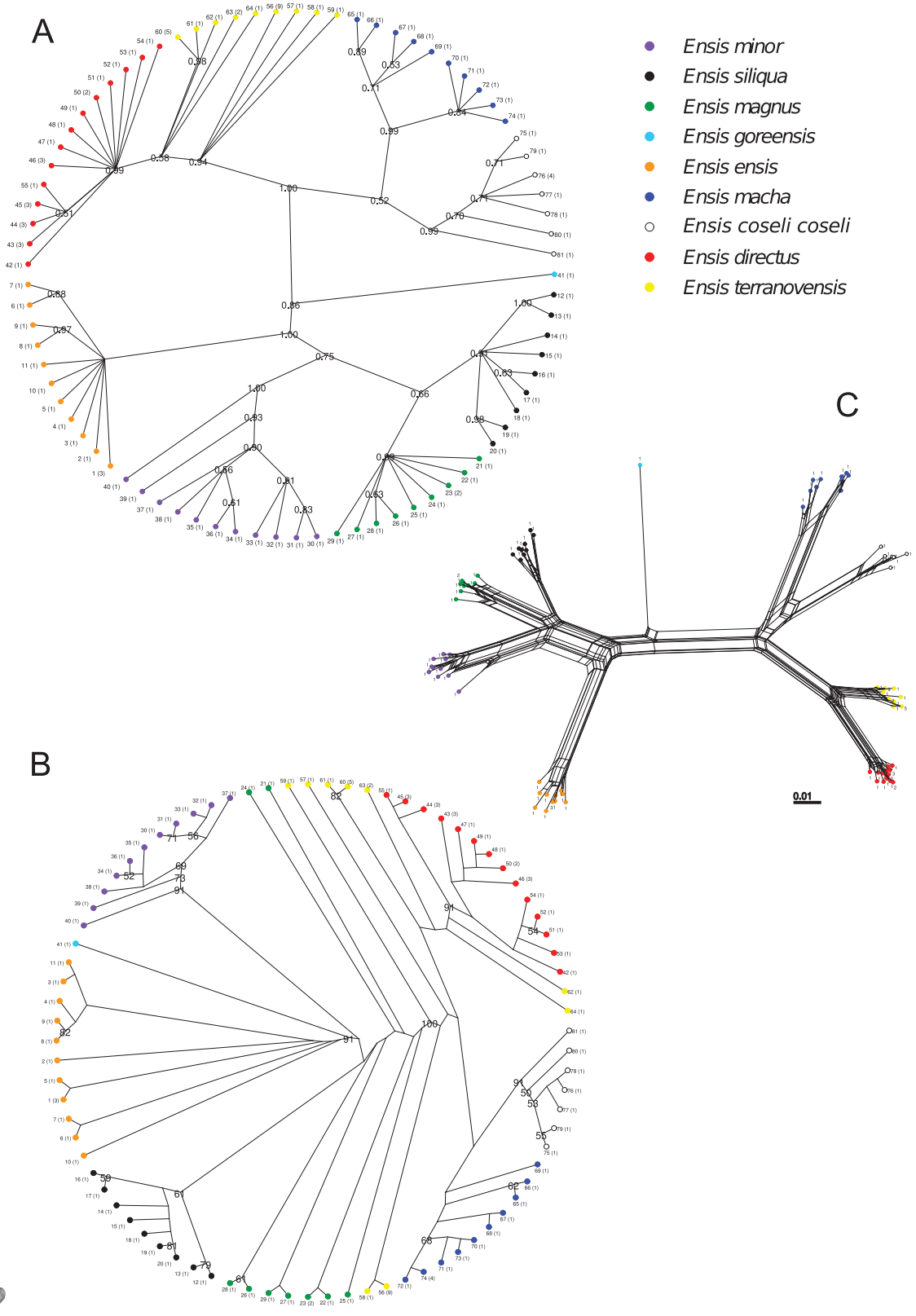


Fig. S2 Phylogenetic relationships among 16S haplotypes. Haplotype frequencies are shown in parentheses in the tree, and at each terminal node, in the network. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from; the white square refers to the haplotype shared by *E. directus* and *E. terranovensis*. **A** Unrooted Bayesian phylogenetic tree. Node confidence values below 0.5 are not shown. **B** Unrooted maximum likelihood phylogenetic tree. Node confidence values below 50 are not shown. **C** Phylogenetic network constructed using the neighbournet algorithm and uncorrected p-distances.

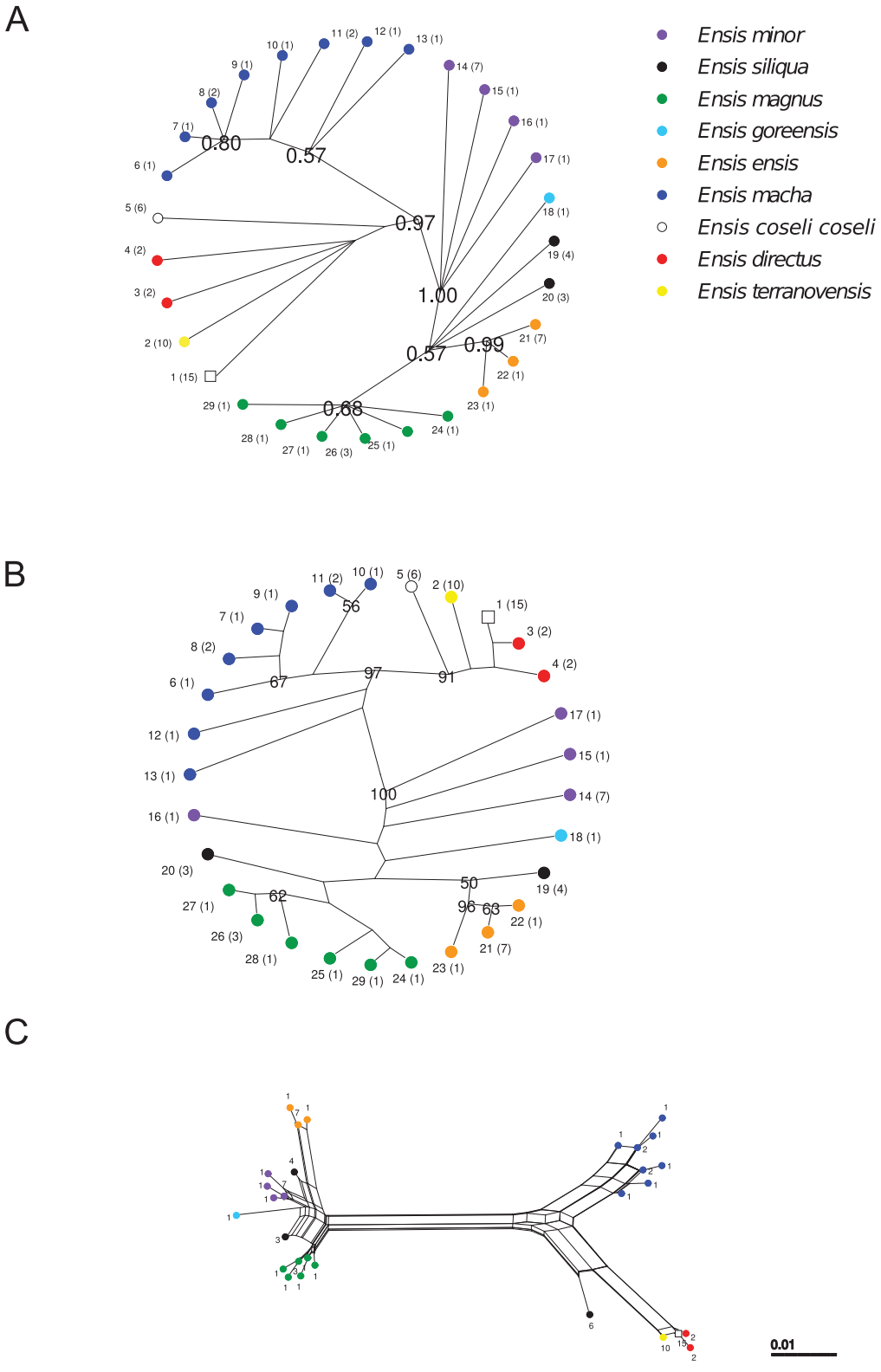


Fig. S3 Unrooted maximum likelihood phylogenetic tree based on the COI+16S haplotype alignment. Haplotype frequencies are shown in parentheses. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from. Node confidence values below 50 are not shown.

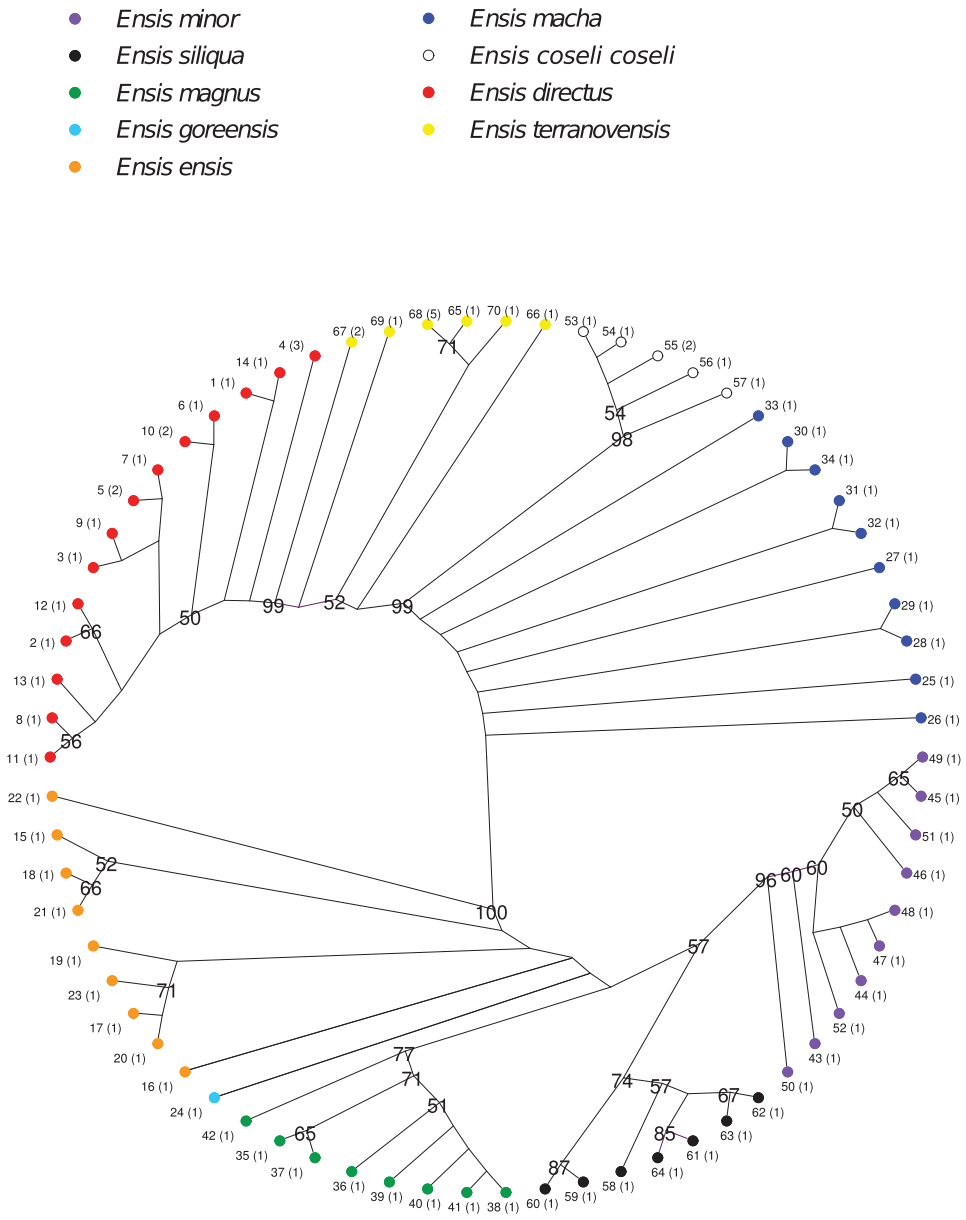


Fig. S4 Phylogenetic relationships among ITS1 sequence-types. Sequence-type frequencies are shown in parentheses in the tree, and at each terminal node, in the network. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from. **A** Unrooted Bayesian phylogenetic tree. Node confidence values below 0.5 are not shown. **B** Unrooted maximum likelihood phylogenetic tree. Node confidence values below 50 are not shown. **C** Phylogenetic network constructed using the neighbournet algorithm and uncorrected p-distances.

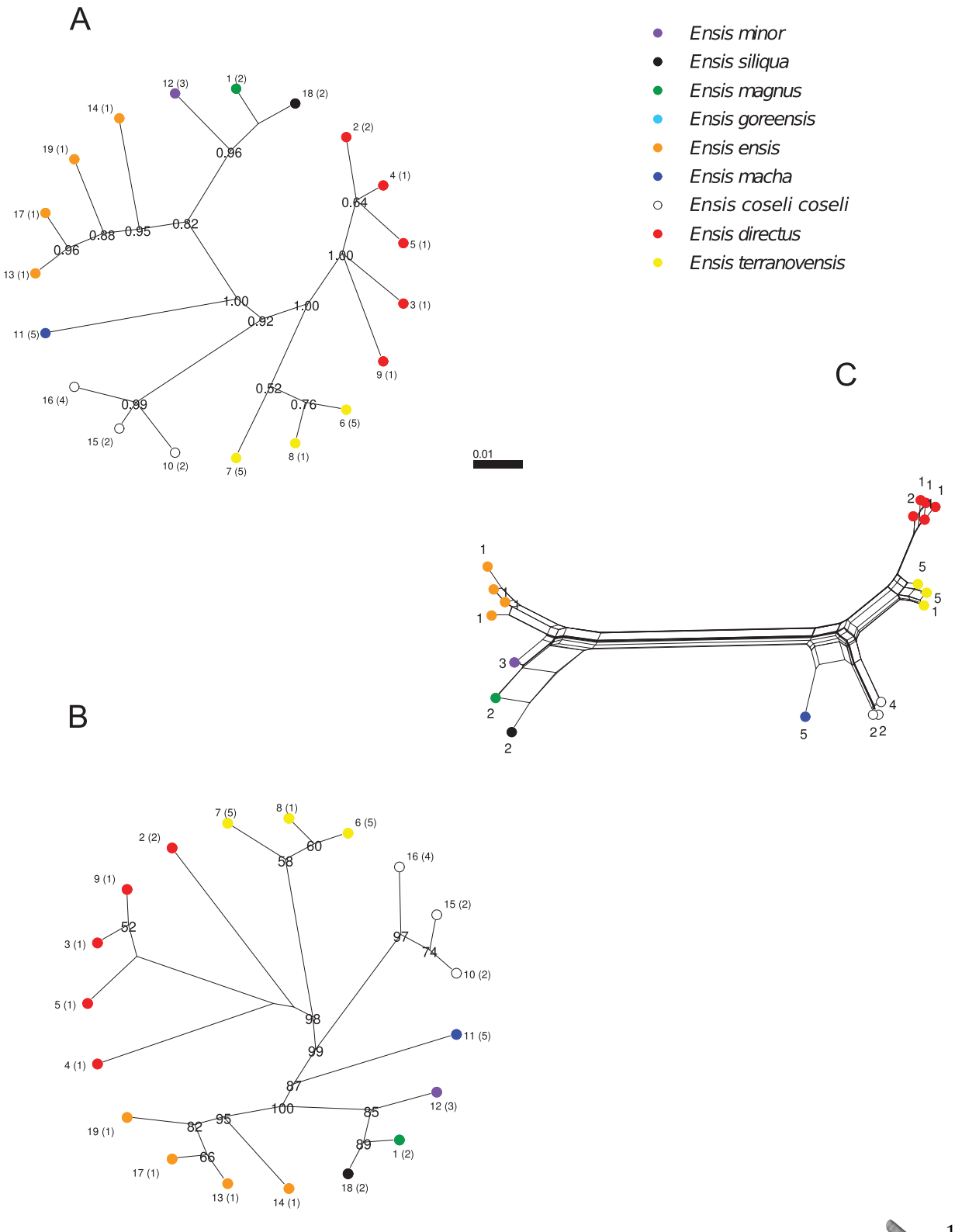


Fig. S5 Phylogenetic relationships among ITS2 sequence-types. Sequence-type frequencies are shown in parentheses in the tree, and at each terminal node, in the network. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from. **A** Unrooted Bayesian phylogenetic tree. Node confidence values below 0.5 are not shown. **B** Unrooted maximum likelihood phylogenetic tree. Node confidence values below 50 are not shown. **C** Phylogenetic network constructed using the neighbournet algorithm and uncorrected p-distances.

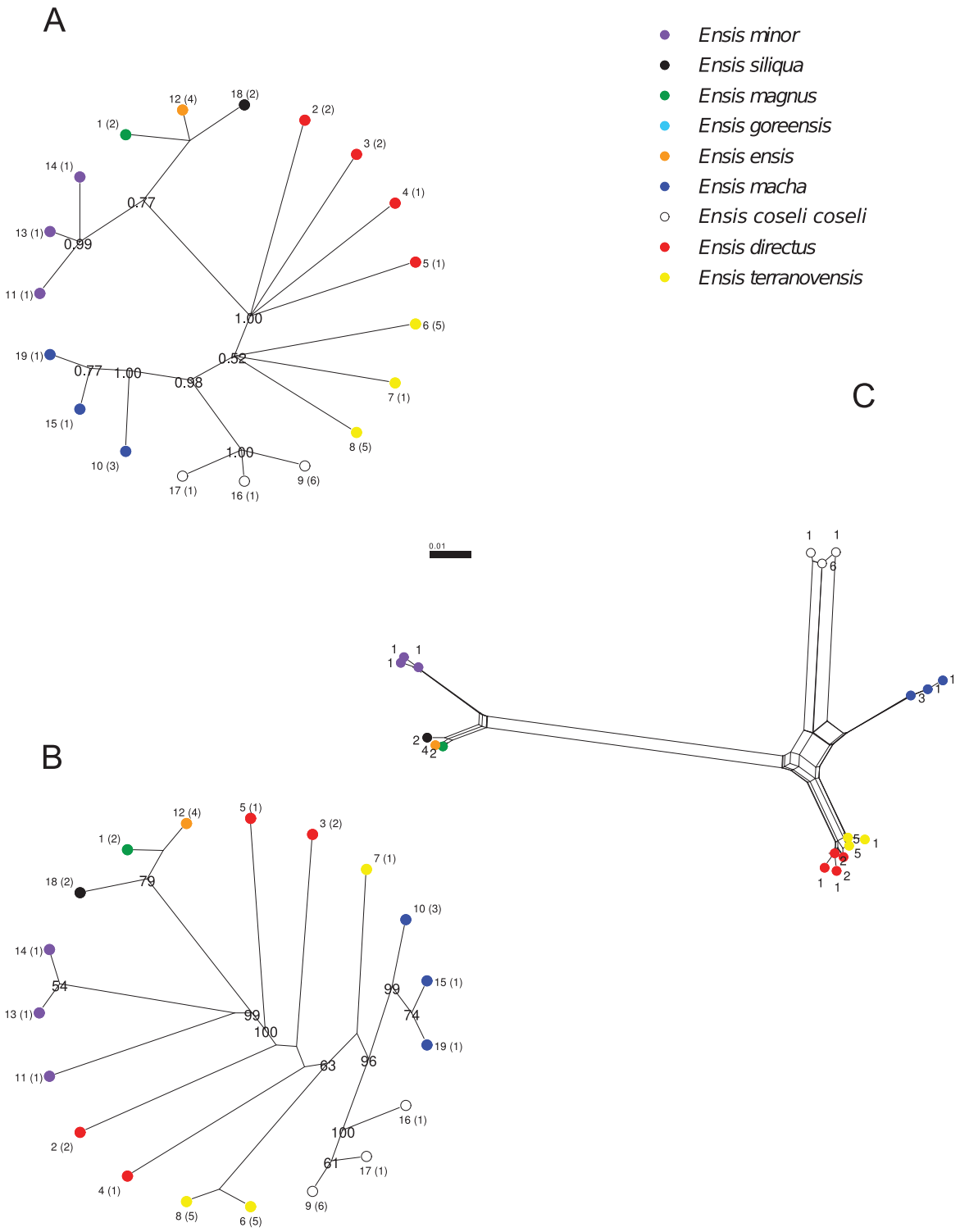


Fig. S6 Unrooted maximum likelihood phylogenetic tree based on the ITS1+ITS2 sequence-type alignment. Sequence-type frequencies are shown in parentheses. Terminal nodes are coloured according to which *Ensis* species the sequence was obtained from. Node confidence values below 50 are not shown.

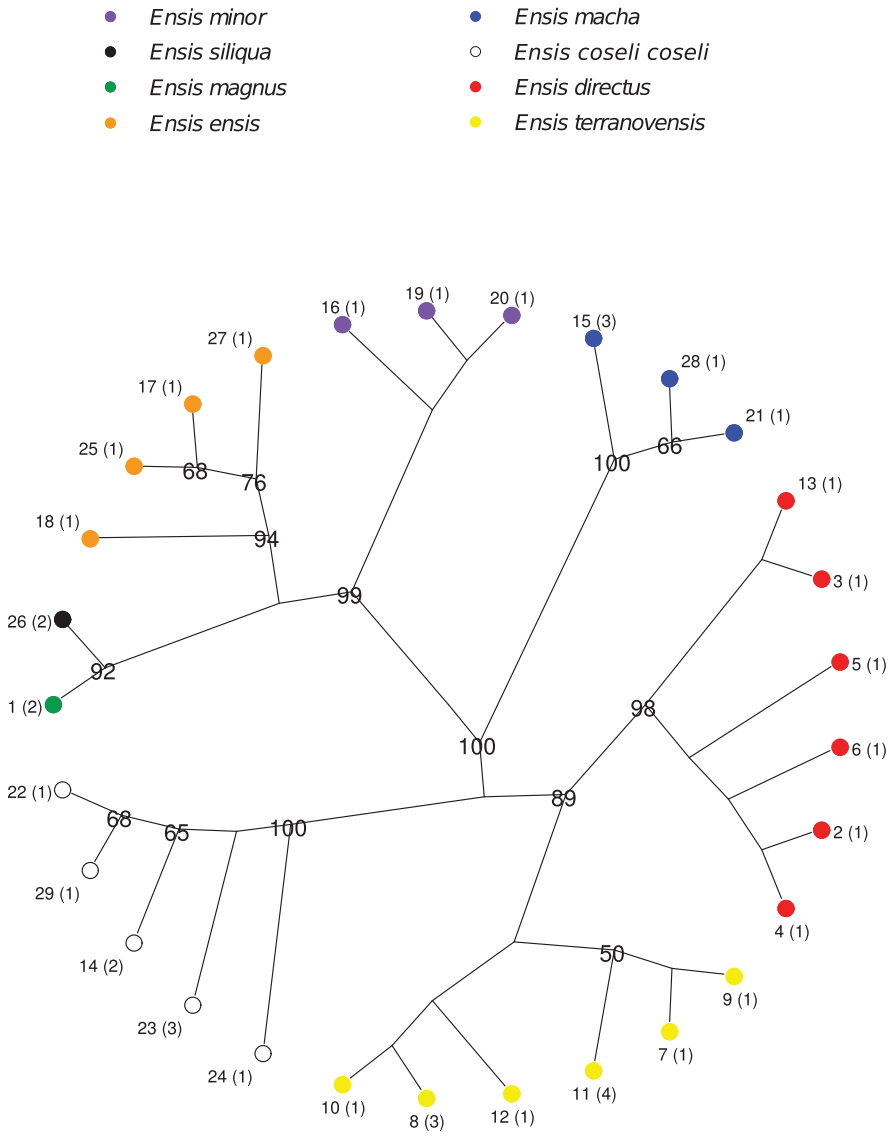


Table S1 Razor shell specimens studied.

Species	Sampling site	Country	Sampled by	Year	Identified by	h-COI	h-16S	h-COI-16S	h-18S	h-5.8S	h-ITS1	h-ITS2	h-ITS1-ITS2	Museum code
Ensis ensis	La Capte	France	E. Egea	2007	R. von Cosel, J. Vierna	7	21	21	6	5	13	12	17	MNHN 40044
Ensis ensis	Borkum reef	Germany	IMARES	2011	J. Cuperus	2				5	19	12	27	RMNH.MOL.303483
Ensis ensis	Borkum reef	Germany	IMARES	2011	J. Cuperus	6	21	18						RMNH.MOL.303484
Ensis ensis	Borkum reef	Germany	IMARES	2011	J. Cuperus	5	21	19						RMNH.MOL.303486
Ensis ensis	Málaga	Spain	L. Kirkendale	2003	L. Kirkendale	10	21	22		5	14	12	18	UF380799
Ensis ensis	52 47 N / 03 33 E	The Netherlands	IMARES	2011	J. Cuperus	9				5	17	12	25	RMNH.MOL.303443
Ensis ensis	52 47 N / 03 33 E	The Netherlands	IMARES	2011	J. Cuperus	3								RMNH.MOL.303441
Ensis ensis	52 47 N / 03 33 E	The Netherlands	IMARES	2011	J. Cuperus	4								RMNH.MOL.303442
Ensis ensis	52 58 N / 03 48 E	The Netherlands	IMARES	2011	J. Cuperus	8	21	15						RMNH.MOL.303437
Ensis ensis	52 58 N / 03 48 E	The Netherlands	IMARES	2011	J. Cuperus	11	21	16						RMNH.MOL.303438
Ensis ensis	52 58 N / 03 48 E	The Netherlands	IMARES	2011	J. Cuperus	1	21	17						RMNH.MOL.303439
Ensis ensis	Central North Sea	The Netherlands	IMARES	2011	J. Cuperus	1	22	20						RMNH.MOL.303480
Ensis ensis	Zeeland	The Netherlands	K. Goudswaard	2011	J. Cuperus	1	23	23						RMNH.MOL.303478
Ensis goreensis	Gorée Bay	Senegal	M. Pin	1987	R. von Cosel	41	18	24						MNHN 17948
Ensis magnus	Borkum reef	Germany	IMARES	2011	J. Cuperus	24	26	39		4	1	1	1	RMNH.MOL.303488
Ensis magnus	Borkum reef	Germany	IMARES	2011	J. Cuperus	26	29	37						RMNH.MOL.303485
Ensis magnus	Borkum reef	Germany	IMARES	2011	J. Cuperus	22	26	38						RMNH.MOL.303487
Ensis magnus	Borkum reef	Germany	IMARES	2011	J. Cuperus	21	27	40						RMNH.MOL.303493
Ensis magnus	Borkum reef	Germany	IMARES	2011	J. Cuperus	23	28	41						RMNH.MOL.303503
Ensis magnus	Ortigueira	Spain	J.M. Casariego	2007	R. von Cosel, J. Vierna	29	25	42	5					MNHN 40043
Ensis magnus	Bonden	Sweden	K. Worsaae, I. Heiner, J. Havenhand	2007	R. von Cosel, J. Vierna	27	26	36	5	5	1	1	1	MNHN 40042
Ensis magnus	Bonden	Sweden	K. Worsaae, I. Heiner, J. Havenhand	2007	R. von Cosel, J. Vierna	28	24	35						MNCN 15.07/12110
Ensis magnus	Bonden	Sweden	K. Worsaae, I. Heiner, J. Havenhand	2007	R. von Cosel, J. Vierna	23								MNCN 15.07/12111
Ensis magnus	52 47 N / 03 33 E	The Netherlands	IMARES	2010	J. Cuperus	25								RMNH.MOL.303440
Ensis minor	La Capte	France	E. Egea	2007	R. von Cosel, J. Vierna	38	14	48		5	12	14	20	RMNH.MOL.307872
Ensis minor	La Capte	France	E. Egea	2007	R. von Cosel, J. Vierna	31	14	49	5	5	12	13	19	MNHN 40045
Ensis minor	Bandol	France	E. Egea	2007	R. von Cosel, J. Vierna	39	14	43						RMNH.MOL.307867
Ensis minor	Bandol	France	E. Egea	2007	R. von Cosel, J. Vierna	35	14	44						RMNH.MOL.307868
Ensis minor	Bandol	France	E. Egea	2007	R. von Cosel, J. Vierna	30	17	45						RMNH.MOL.307869
Ensis minor	Bandol	France	E. Egea	2007	R. von Cosel, J. Vierna	33	16	46						RMNH.MOL.307870
Ensis minor	Bandol	France	E. Egea	2007	R. von Cosel, J. Vierna	36	14	47						RMNH.MOL.307871
Ensis minor	Ría de Vigo	Spain	A. Martínez	2007	R. von Cosel, J. Vierna	37	14	52	5	5	12	11	16	MNHN 40047
Ensis minor	Lira	Spain	P. Pita, D. Fernández	2008	J. Cuperus	40	14	50						RMNH.MOL.303495
Ensis minor	Lira	Spain	P. Pita, D. Fernández	2008	J. Cuperus	34								RMNH.MOL.303497
Ensis minor	Ría de Vigo	Spain	A. Martínez	2007	R. von Cosel, J. Vierna	32	15	51	5					MNHN 40046
Ensis siliqua	Borkum reef	Germany	IMARES	2011	J. Cuperus	18	20	58						RMNH.MOL.303490
Ensis siliqua	Borkum reef	Germany	IMARES	2011	J. Cuperus	12	20	59						RMNH.MOL.303492
Ensis siliqua	Borkum reef	Germany	IMARES	2011	J. Cuperus	13	20	60						RMNH.MOL.303494
Ensis siliqua	Cedeira	Spain	J. Fernández	2010	J. Cuperus	19	19	61		5	18	18	26	RMNH.MOL.303499
Ensis siliqua	Cedeira	Spain	J. Fernández	2010	J. Cuperus	16			5					RMNH.MOL.303498
Ensis siliqua	Cedeira	Spain	J. Fernández	2010	J. Cuperus	17								RMNH.MOL.303500
Ensis siliqua	Cedeira	Spain	J. Fernández	2010	J. Cuperus	14	19	62						RMNH.MOL.303501
Ensis siliqua	Cedeira	Spain	J. Fernández	2010	J. Cuperus	15	19	63						RMNH.MOL.303502
Ensis siliqua	Central North Sea	The Netherlands	IMARES	2011	J. Cuperus	20	19	64		5	18	18	26	RMNH.MOL.303481
Ensis directus	Fiskebäckskil	Sweden	J. Vierna	2007	J. Vierna	50	3	5	2					MNHN 40050
Ensis directus	Bergen / Egmond	The Netherlands	IMARES	2010	J. Cuperus	45	1	1						RMNH.MOL.303465
Ensis directus	Bergen / Egmond	The Netherlands	IMARES	2010	J. Cuperus	43	4	2						RMNH.MOL.303466
Ensis directus	Bergen / Egmond	The Netherlands	IMARES	2010	J. Cuperus	49	1	3						RMNH.MOL.303468
Ensis directus	Bergen / Egmond	The Netherlands	IMARES	2010	J. Cuperus	44	1	4						RMNH.MOL.303469
Ensis directus	Wadden Sea	The Netherlands	IMARES	2008	J. Cuperus	47	1	7						RMNH.MOL.303351
Ensis directus	Wadden Sea	The Netherlands	IMARES	2008	J. Cuperus	51	1	8						RMNH.MOL.303352
Ensis directus	Wadden Sea	The Netherlands	IMARES	2008	J. Cuperus	53	1	9						RMNH.MOL.303353
Ensis directus	Wadden Sea	The Netherlands	IMARES	2008	J. Cuperus	46	1	10						RMNH.MOL.303354
Ensis directus	Wadden Sea	The Netherlands	IMARES	2008	J. Cuperus	50	3	5						RMNH.MOL.303355
Ensis directus	Wadden Sea	The Netherlands	IMARES	2008	J. Cuperus	44	1	4						RMNH.MOL.303356

Species	Sampling site	Country	Sampled by	Year	Identified by	h-COI	h-16S	h-COI-16S	h-18S	h-5.8S	h-ITS1	h-ITS2	h-ITS1-ITS2	Museum code
Ensis directus	Zeeland	The Netherlands	Meromar	2010	J. Cuperus	46	1	10						RMNH.MOL.30339
Ensis directus	Zeeland	The Netherlands	Meromar	2010	J. Cuperus	52	1	11						RMNH.MOL.303394
Ensis directus	Zeeland	The Netherlands	Meromar	2010	J. Cuperus	42	4	12						RMNH.MOL.303396
Ensis directus	Zeeland	The Netherlands	Meromar	2010	J. Cuperus	44	1	4						RMNH.MOL.303397
Ensis directus	The Wash	United Kingdom	D. Palmer	2007	R. von Cosel, J. Vierna	43	1	6		3	9	2	13	MNCN 15.07/12101
Ensis directus	Jacksonville	United States	M. Bemis, J. Moore	2010	J. Vierna	54	1	13						UF446939
Ensis directus	Jacksonville	United States	M. Bemis, J. Moore	2010	J. Vierna	55	1	14						UF446941
Ensis directus	Cobscook Bay	United States	T. Sheehan (Gulf of Maine)	2008	J. Vierna	45				2	2	2	2	MNCN 15.07/11734
Ensis directus	Cobscook Bay	United States	T. Sheehan (Gulf of Maine)	2008	J. Vierna	45				2	2	4	4	MNCN 15.07/11734
Ensis directus	Cobscook Bay	United States	T. Sheehan (Gulf of Maine)	2008	J. Vierna	46				2	4	3	5	MNCN 15.07/11734
Ensis directus	Cobscook Bay	United States	T. Sheehan (Gulf of Maine)	2008	J. Vierna	48				2	3	3	3	MNCN 15.07/11734
Ensis directus	Cobscook Bay	United States	T. Sheehan (Gulf of Maine)	2008	J. Vierna	43				2	5	5	6	MNCN 15.07/11734
Ensis macha	Puerto Lobos	Argentina	L. Orensanz	2009	J. Vierna	67	10	32						MNCN 15.07/12108
Ensis macha	Puerto Lobos	Argentina	L. Orensanz	2009	J. Vierna	69	11	33						MNCN 15.07/12109
Ensis macha	Puerto Lobos	Argentina	L. Orensanz	2009	J. Vierna	65	13	30		2	11	10	15	MNCN 15.07/12106
Ensis macha	Puerto Lobos	Argentina	L. Orensanz	2009	J. Vierna	68	11	31		2	11	10	15	MNCN 15.07/12107
Ensis macha	Puerto Lobos	Argentina	L. Orensanz	2009	J. Vierna	66	12	34	1	2	11	10	15	MNHN IM-2009-8446
Ensis macha	Playa Dichato	Chile	A. de la Torriente	2007	R. von Cosel, J. Vierna	70	6	25						MNCN 15.07/12102
Ensis macha	Playa Dichato	Chile	A. de la Torriente	2007	R. von Cosel, J. Vierna	74	8	27						MNCN 15.07/12103
Ensis macha	Playa Dichato	Chile	A. de la Torriente	2007	R. von Cosel, J. Vierna	71	7	28						MNCN 15.07/12104
Ensis macha	Playa Dichato	Chile	A. de la Torriente	2007	R. von Cosel, J. Vierna	73	9	26	1	2	11	19	28	MNHN 40048
Ensis macha	Playa Dichato	Chile	A. de la Torriente	2007	R. von Cosel, J. Vierna	72	8	29		2	11	15	21	MNCN 15.07/12105
Ensis coseli coseli	99th Street *	United States	A. Schulze	2010	J. Vierna	77	5	53						MNCN 15.07/12112
Ensis coseli coseli	Jamaica Beach *	United States	A. Schulze	2010	J. Vierna	76	5	55						MNCN 15.07/12115
Ensis coseli coseli	Christmas Bay	United States	A. Schulze	2009	J. Vierna	76	5	55	4	2	10	9	14	MNHN IM-2009-8447
Ensis coseli coseli	Jamaica Beach *	United States	A. Schulze	2010	J. Vierna	76				2	10	9	14	MNCN 15.07/12174
Ensis coseli coseli	Levy County	United States	A. Anker et al.	2008	J. Vierna	80	5	57	4	2	15	16	22	UF424378
Ensis coseli coseli	Levy County	United States	A. Anker et al.	2008	J. Vierna	79				2	16	9	23	UF424379
Ensis coseli coseli	99th Street *	United States	A. Schulze	2010	J. Vierna	78	5	54		2	16	9	23	MNCN 15.07/12113
Ensis coseli coseli	Jamaica Beach *	United States	A. Schulze	2010	J. Vierna	75	5	56		2	15	9	29	MNCN 15.07/12114
Ensis coseli coseli	Jamaica Beach *	United States	A. Schulze	2010	J. Vierna	76				2	16	9	23	MNCN 15.07/12173
Ensis coseli coseli	Levy County	United States	A. Anker et al.	2008	J. Vierna	81				2	16	17	24	UF424380
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	60				2	7	6	8	MNCN 15.07/15001
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	62				2	7	6	8	MNCN 15.07/15002
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	59				2	7	7	10	MNCN 15.07/15002
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	60				1	7	8	12	MNCN 15.07/15002
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	56				2	7	6	8	MNHN IM-2009-16708
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	56	1	65	3	2	6	6	7	MNHN 40049
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	60				2	6	8	11	MNCN 15.07/15002
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	58				2	6	8	11	CMNML 096168
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	56				2	6	8	11	CMNML 096168
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	56				2	6	8	11	MNHN IM-2009-16706
Ensis terranovensis	Long Pond	Canada	P. Sargent, R. O'Donnell	2007	J. Vierna	61				2	8	6	9	MNCN 15.07/15002
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	63	2	66						RMNH.MOL.303511
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	60	2	67						RMNH.MOL.303512
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	56	2	68						RMNH.MOL.303513
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	56	2	68						RMNH.MOL.303514
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	64	2	69						RMNH.MOL.303515
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	60	2	67						RMNH.MOL.303516
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	56	2	68						RMNH.MOL.303517
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	63	2	70						RMNH.MOL.303518
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	56	2	68						RMNH.MOL.303520
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	56	2	68						RMNH.MOL.303521
Ensis terranovensis	Long Pond	Canada	R.J. Thompson	2011	J. Cuperus	57								RMNH.MOL.303522
						109	**80**	**80**	**13**	**41**	**41**	**41**	**41**	

h- stands for the haplotype or sequence-type number. MNHN, Muséum national d'Histoire naturelle (Paris, France); RMNH, Naturalis Biodiversity Center (Leiden, The Netherlands); UF, Florida Museum of Natural History (Gainesville, Florida, USA); MNCN, Museo Nacional de Ciencias Naturales (Madrid, Spain); CMNML, Canadian Museum of Nature (Ottawa, Canada); * Galveston Bay.

Table S2 Accession numbers of sequences retrieved from Vierna *et al*. (2012).

COI	ITS1	ITS2
HE661776	HE661883	HE662004
HE661664	HE661945	HE662011
HE661715	HE661946	HE662095
HE661729	HE661947	HE662096
HE661730	HE661948	HE662097
HE661748	HE661949	HE662098
HE661749	HE661984	HE662128
HE661765	HE661985	HE662129
HE661766	HE661986	HE662130
HE661767	HE661987	HE662132
HE661768	HE661988	HE662133
HE661771	HE661989	HE662134
HE661772	HE661990	HE662135
HE661773	HE661991	HE662136
HE661774	HE661992	HE662137
HE661775	HE661993	HE662140
HE661777	HE661997	HE662142

Table S3 Models of evolution for each partition.

COI	HKY+I+G
16S	GTR+I+G
ITS1	HKY+G
ITS2	GTR+I

Table S4 Mean interspecific and mean intraspecific divergence based on the COI alignment.

	Ensis ensis	*Ensis goreensis*	*Ensis magnus*	*Ensis minor*	*Ensis siliqua*	*Ensis directus*	*Ensis macha*	*Ensis coseli coseli*	*Ensis terranovensis*	**d**	**e**
Ensis ensis	-	0.019	0.017	0.017	0.017	0.021	0.020	0.020	0.021	0.015	0.003
Ensis goreensis	0.143	-	0.018	0.019	0.018	0.020	0.020	0.020	0.020	n/a	n/a
Ensis magnus	0.115	0.130	-	0.013	0.012	0.021	0.020	0.021	0.020	0.011	0.003
Ensis minor	0.124	0.141	0.077	-	0.013	0.022	0.020	0.020	0.021	0.021	0.005
Ensis siliqua	0.116	0.130	0.064	0.078	-	0.021	0.019	0.020	0.019	0.019	0.004
Ensis directus	0.179	0.168	0.178	0.195	0.176	-	0.019	0.018	0.012	0.008	0.002
Ensis macha	0.175	0.162	0.179	0.180	0.171	0.134	-	0.016	0.017	0.026	0.006
Ensis coseli coseli	0.169	0.167	0.186	0.171	0.165	0.123	0.104	-	0.017	0.013	0.003
Ensis terranovensis	0.176	0.168	0.163	0.178	0.154	0.058	0.114	0.108	-	0.008	0.003

Below diagonal, mean uncorrected p-distance between species pairs; above diagonal, standard errors. d, mean uncorrected p-distance within species; e, standard errors. n/a, not applicable.

Table S5 Mean interspecific and mean intraspecific divergence based on the 16S alignment.

	Ensis ensis	*Ensis goreensis*	*Ensis magnus*	*Ensis minor*	*Ensis siliqua*	*Ensis directus*	*Ensis macha*	*Ensis coseli coseli*	*Ensis terranovensis*	d	e
Ensis ensis	-	0.001	0.001	0.001	0.009	0.072	0.075	0.075	0.072	0.007	0.007
Ensis goreensis	0.032	-	0.001	0.009	0.009	0.072	0.075	0.075	0.072	48n	48n
Ensis magnus	0.032	0.032	-	0.00/	0.00a	0.075	0.073	0.073	0.075	0.00a	0.003
Ensis minor	0.033	0.079	0.076	-	0.00a	0.075	0.073	0.073	0.075	0.003	0.007
Ensis siliqua	0.072	0.079	0.077	0.077	-	0.075	0.075	0.073	0.075	0.003	0.003
Ensis directus	0.0/ /	0.0/ 9	0.096	0.0/ 3	0.0/ 0	-	0.070	0.001	0.005	0.007	0.007
Ensis macha	0.091	0.09a	0.090	0.0a/	0.09a	0.020	-	0.070	0.077	0.001	0.005
Ensis coseli coseli	0.093	0.09a	0.0a9	0.0a/	0.097	0.03/	0.05/	-	0.001	0.000	0.000
Ensis terranovensis	0.0/ 2	0.0/ 6	0.099	0.096	0.09/	0.005	0.023	0.032	-	0.007	0.007

Below dingo4nl, men4 u4corrected p-distn4ce betwee4 species pnirs; nbove dingo4nl, stn4dnrd errors. d, men4 u4corrected p-distn4ce withi4 species; e, stn4dnrd errors. 48n, 4ot npplicnble.

Table S6 Mean interspecific and mean intraspecific divergence based on the ITS1 alignment.

	Ensis ensis	*Ensis g or nms*	*Ensis g inau*	*Ensis siliqmo*	*Ensis diuectms*	*Ensis g ocho*	*Ensis caseli caseli*	*Ensis teuonavensis*	d	e
Ensis ensis	-	0.010	0.008	0.012	0.018	0.016	0.017	0.017	0.009	0.004
Ensis g or nms	0.035	-	0.006	0.007	0.018	0.016	0.017	0.017	0.000	0.000
Ensis g inau	0.025	0.011	-	0.009	0.018	0.016	0.017	0.017	0.000	0.000
Ensis siliqmo	0.049	0.014	0.025	-	0.018	0.016	0.017	0.017	0.000	0.000
Ensis diuectms	0.104	0.107	0.100	0.107	-	0.013	0.012	0.008	0.006	0.003
Ensis g ocho	0.087	0.084	0.077	0.084	0.054	-	0.010	0.012	0.000	0.000
Ensis caseli caseli	0.094	0.094	0.088	0.094	0.049	0.030	-	0.010	0.003	0.002
Ensis teuonavensis	0.095	0.098	0.091	0.098	0.022	0.044	0.035	-	0.003	0.002

Below diagonal, mean uncorrected p-distance between species pairs; above diagonal, standard errors. d, mean uncorrected p-distance within species; e, standard errors.

Table S7 Mean interspecific and mean intraspecific divergence based on the ITS2 alignment.

	Ensis ensis	*Ensis g or nms*	*Ensis g inau*	*Ensis siliqmo*	*Ensis diuectms*	*Ensis g ocho*	*Ensis caseli caseli*	*Ensis teuonavensis*	**d**	**e**
Ensis ensis	-	0.001	0.082	0.006	0.028	0.022	0.027	0.028	0.000	0.000
Ensis g or nms	0.001	-	0.082	0.006	0.020	0.022	0.022	0.020	0.000	0.000
Ensis g inau	0.079	0.079	-	0.082	0.028	0.022	0.027	0.028	0.006	0.001
Ensis siliqmo	0.004	0.004	0.079	-	0.028	0.022	0.027	0.028	0.000	0.000
Ensis diuectms	0.883	0.881	0.824	0.885	-	0.081	0.089	0.003	0.003	0.007
Ensis g ocho	0.871	0.871	0.875	0.871	0.039	-	0.086	0.081	0.001	0.007
Ensis caseli caseli	0.818	0.875	0.837	0.813	0.090	0.057	-	0.085	0.002	0.002
Ensis teuonavensis	0.881	0.882	0.829	0.886	0.009	0.037	0.051	-	0.007	0.002

Below diagonal, mean uncorrected p-distance between species pairs; above diagonal, standard errors. d, mean uncorrected p-distance within species; e, standard errors.

5 CORRIGENDUM

5. CORRIGENDUM

Since the publication of the scientific articles that comprise this thesis, we have noticed some errors that, in general, do not alter the main conclusions of the articles but should nevertheless be corrected. Therefore, we list them below. The typographical errors are not recorded, unless they were misleading.

Corrections to CHAPTER 4.1.1: **Joaquín Vierna, K Thomas Jensen, Andrés Martínez-Lage, Ana M González-Tizón (2011) The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae). Heredity 107:127-142.**

- During the revision of the proofs of this article, the parentheses of some species authorities were incorrectly modified by the staff of the editorial office. The correct authorities are recorded in CHAPTER 2.
- The sequences supposedly obtained from *E. siliqua* (Linné, 1758) specimens belong to *E. minor* (Chenu, 1843) individuals. This has been clarified in CHAPTER 4.4 and the DDBJ/EMBL/GenBank records have been updated accordingly.

Corrections to CHAPTER 4.1.2: **Joaquín Vierna, Stefanie Wehner, Christian Höner zu Siederdissen, Andrés Martínez-Lage, Manja Marz (2013) Systematic analysis and evolution of 5S ribosomal DNA in metazoans. Heredity 111:410-421.**

- In the 'Orthologous 5S rRNA genes' subsection, we stated that (1) vertebrate 5S rRNA sequences are clearly evolutionary separated from other metazoan sequences, and (2) basal deuterostomes (Hemichordata, Tunicata and Cephalochordata) and nematodes share high sequence similarity, whereas the sequences of other metazoans (Arthropoda, Lophotrochozoa, Cnidaria, Porifera and Placozoa) clustered into a distinct 5S rRNA group. Whereas the first statement is correct (see network, Figure 2 left), the second one should be revised. Basal Deuterostomia and Nematoda do not cluster into a distinct group, nor the sequences of other Metazoa do. In fact, only the sequences of the Vertebrata clustered apart from all others; and those obtained from basal Deuterostomia, Nematoda, and other Metazoa do not show a clear clustering pattern by phylogenetic group. This also applies to Figure 2 caption.

Corrections to CHAPTER 4.3: **Joaquín Vierna, K. Thomas Jensen, Ana M. González-Tizón, Andrés Martínez-Lage (2012) Population genetic analysis of *Ensis directus* unveils high genetic variation in the introduced range and reveals a new species from the NW Atlantic. Marine Biology 159:2209-2227.**

- During the revision of the proofs of this article, the staff of the editorial office changed some of the references by Cosel von R to von Cosel R. Thus, these references are located at different positions on the references list.

Corrections to CHAPTER 4.4: **Joaquín Vierna, Joël Cuperus, Andrés Martínez-Lage, Jeroen M. Jansen, Alejandra Perina, Hilde Van Pelt, Ana M. González-Tizón (2013) Species delimitation and DNA barcoding of Atlantic *Ensis* (Bivalvia, Pharidae). Zoologica Scripta doi:10.1111/zsc.12038.**

- As already stated in CHAPTER 2.2, in this article we introduced the names *E. coseli coseli* Vierna, 2013 and *E. coseli megistus* Pilsbry and McGinty, 1943, but these names resulted to be non-code-compliant according to articles 23.3.1, 46.1, and 47.2 of the International Code of Zoological Nomenclature. The correct names are *E. megistus coseli* Vierna, 2013 and *E. megistus megistus* Pilsbry and McGinty, 1943.

Corrections to CHAPTER 10.1: **Joaquín Vierna, Ana M. González-Tizón, Andrés Martínez-Lage (2009) Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). Biochemical Genetics 47:635–644.**

- The parentheses of some taxa authorities were incorrectly written. The correct authorities are recorded in CHAPTER 2.
- Sequences FM201453 and FM211693 that were apparently obtained from '*Ensis arcuatus*' (*E. magnus* Schumacher, 1817) are probably the result of a cross-contamination. Therefore, we have decided to delete these records from DDBJ/EMBL/GenBank.
- In the 'Introduction' section there is a mistake in the sentence *these genomic regions has been widely used in molecular evolutionary studies* (…), that should be read as follows:

these genomic regions have been widely used in molecular evolutionary studies (…).

– In the 'Discussion' section there is a mistake in the sentence *this indicates that they split long ago, and probably were already present in the genome of the most common ancestor of these species (ancestral polymorphism)*, that should be read as follows: *this indicates that they split long ago, and probably were already present in the genome of the most recent common ancestor of these species (ancestral polymorphism).*

Corrections to CHAPTER 10.2: **Joaquín Vierna, Andrés Martínez-Lage, Ana M. González-Tizón, (2010) Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. Genome 53:23-34.**

– The parentheses of some taxa authorities were incorrectly written. The correct authorities are recorded in CHAPTER 2.

– Maximum likelihood phylogenetic trees are not “majority-rule consensus trees”, as stated on the captions of Figures 2 and 3. Rather, they are the maximum likelihood trees inferred by the software PhyML.

– Upon the publication of CHAPTER 4.3 it became clear that the *Ensis* specimens from Long Pond belong to a new species (*E. terranovensis* Vierna & Martínez-Lage, 2012). Therefore, one of the main conclusions of the article, the applicability of the ITS1 and ITS2 regions to differentiate among *E. directus* individuals from different geographic areas, cannot be maintained.

6 GENERAL DISCUSSION

6. GENERAL DISCUSSION

6.1 Contributions of this thesis to the understanding of the evolution of 5S rDNA

Until 1990, most multigene families were thought to be subjected to concerted evolution (Arnheim et al. 1980; Nei and Rooney 2005), being the nuclear rRNA genes the archetypal example (see Eickbush and Eickbush 2007). Under the concerted evolution model, all members of the multigene family are assumed to evolve as a unit in concert, and a mutation occurring in a repeat spreads through the entire array by repeated occurrence of unequal crossover or gene conversion (Nei and Rooney 2005).

After the early 1990s researchers began to unveil higher levels of intraspecific nucleotide diversity within certain multigene families that were not consistent with the concerted evolution model. These findings, together with other atypical features observed across multigene family members (as the presence of between-species clustering patterns in phylogenies; the presence of pseudogenes) motivated the proposal of a new evolutionary model (Eirín-López et al. 2012).

According to this model, which was named 'evolution by the birth-and-death process' (Nei and Hughes 1992), new genes are created by gene duplication and some duplicated genes are maintained in the genome for a long time, whereas others are deleted or inactivated through deleterious mutations (Nei and Rooney 2005). Birth-and-death evolution is characterised not only by the presence of different types of gene sequences, which may or may not be shared among closely related species, but also by the frequent occurrence of pseudogenised gene copies.

Interestingly, it was not until 2004 that a report was published on the birth-and-death evolution of a rRNA gene family, the 18S rRNA (Rooney 2004). Since then, several other studies (e.g. Rooney and Ward 2005; Fujiwara et al. 2009; Vierna et al. 2009; Vizoso et al. 2011; Perina et al. 2011; Pinhal et al. 2011) have been published supporting the evolution of 5S rDNA according to mixed models (birth-and-death evolution with strong purifying selection; mixed process of concerted and birth-and-death evolution, see Nei and Rooney 2005) in which the birth-and-death evolution play a substantial role. In fact, birth-and-death evolution is slowly starting to replace concerted evolution as the 'default' model to explain the long-term evolution of multigene families (see Eirín-López et al. 2012).

Some of the results of this thesis (CHAPTERS 4.1.1. and 4.1.2) are in line with these recent findings and support the idea that the evolution of 5S rDNA in metazoans is a complex issue driven by various evolutionary processes in addition to the homogenising mechanisms typically involved in

concerted evolution.

In CHAPTER 4.1.1, we characterised different types of 5S rDNA repeats in detail, and studied the evolution of this multigene family, within the family Pharidae. This allowed us to confirm that the high levels of intragenomic divergence of the *Ensis* NTS region (Vierna et al. 2009) are shared by other *Ensis* species, and by other Pharidae; moreover, pseudogenes, which are an evidence of birth-and-death evolution, that were not found in previous surveys (Vierna et al. 2009), were reported in this chapter (putative pseudogenes and truncated copies); finally, the intermixed organisation of divergent NTS sequences, previously reported for some *Ensis* (Vierna et al. 2009; Fernández-Tajes and Méndez 2009), was found in other Pharidae genera (*S. patula* and *E. cultellus*). Thus, our results confirm that, in these razor shell species, 5S rDNA evolution is not consistent with a simple concerted evolution scenario, but with an evolutionary scenario driven by birth-and-death processes, purifying selection, and the homogenising mechanisms involved in concerted evolution.

In CHAPTER 4.1.2, we climbed up several taxonomic levels, and demonstrated that various features that had previously been observed within the 5S rDNA region for one or a few species, are widespread in metazoans. For example, intragenomic differences in size and base composition within NTS regions, were found in several metazoan lineages, consistently with previous studies (e.g. Gornung et al. 2007; Perina et al. 2011); similarly, different types of putatively functional 5S rRNA coding regions were found to occur within genomes, as in Perina et al. (2011). The internal promoters, potential upstream regulatory regions, and potential termination signals characterised were congruent with previous findings, but new conserved motifs were also reported. Interestingly, some features remain unclear, for example, why the supposedly conserved upstream TATA-like box was only sampled in 13 out of 97 metazoan genomes. The same applies to the mammalian D-box, as eight nucleotides (out of 12) were found to be conserved in 25 mammals, but they were not sampled in eight other mammal species. In the same way, the supposedly conserved TTTT signal was found in 40 species, but not in the remaining 57. Therefore, this opens up possibilities for further research, because some essential questions (are the 5S rRNA coding regions that lack the TATA-like upstream box, or the TTTT termination signal, correctly transcribed?) remain unanswered.

Peterson et al. (1980) demonstrated the existence of two different types of 5S rRNA molecules in *Xenopus*, one being transcribed in the oocyte, and the other one, in somatic tissues. Dimarco et al. (2012) reported another case in which different 5S rRNA variants were expressed in the oocyte and in four embryonic stages in the sea urchin *Paracentrotus lividus*. They found five different variants being expressed in all five tissues studied. Four of these variants were associated, at the genomic

level, to a 700 bp NTS cluster, and one other variant was associated to 900 bp and 950 bp NTS clusters, described in a previous report (Caradonna et al. 2007). When analysing *P. lividus* adult tissues, Bellavia et al. (2013) found out that the four variants associated to the 700 bp NTS cluster were not expressed and are, therefore, embryo-specific. In contrast, the variant transcribed from the 900 bp and 950 bp clusters was present in all developmental stages, including adult tissues, in which it was the only existing variant. Remarkably, these studies support that at least some of the paralog 5S rRNA coding regions that we found in CHAPTERS 4.1.1 and 4.1.2 can be functional variants being expressed in the same or different tissues. Therefore, now that 5S rDNA repeats are characterised in dozens of metazoan species, and that transcriptome sequencing experiments (RNA-seq) are getting more affordable, it may be the time to determine what 5S rRNA variants are being transcribed, along with their genomic context (upstream and downstream elements) in metazoan species. Indeed, this is an interesting topic for further research.

In CHAPTER 4.1.1 we reported the occurrence of linked units of 5S rDNA and U1 snDNA in ten Pharidae species from four different genera (*Ensis, Ensiculus, Pharus,* and *Siliqua*). Based on the occurrence of such a linkage in all species tested (except for *E. goreensis*, probably due to the poor quality of the extracted DNA from the museum specimen), and on the alignment of the IGS region, we concluded that the most parsimonious explanation was that the linkage happened only once, in a common ancestor to all Pharidae species studied.

In a recent study, Beauparlant and Drouin (2013) investigated the linkage of 5S rDNA and spliced-leader (SL) genes in *Trypanosoma*, a group of parasitic flagellate protozoa. The linkages were observed in 11 out of 17 species, however they concluded that they were not acquired from a common ancestor but were the results of independent insertions of 5S rDNA within the SL array. Therefore, even though their data resemble that obtained in razor shells, they came to conclusions that slighly differ from ours.

Based mainly on the analysis of the flanking regions of inserted 5S rRNA coding regions, Beauparlant and Drouin (2013) concluded that the insertions were more likely to have happened independently. By examining the upstream regions of NTS sequences, the authors concluded that some 5S rRNA coding regions (those showing a poly-A tail) had been transposed through an RNA intermediate, whereas some others showed no evidence of retroposition and should had been inserted by means of unequal crossing-over or extrachromosomal covalently closed circular DNAs. Moreover, the authors claimed that the internal promoters of 5S rRNA coding regions allowed these coding regions to be transposed without any selective disadvantage. However, the upstream conserved region, that is essential for efficient transcription in *Caenorhabditis elegans* and *C.*

briggsae (Nelson et al. 1998), *Neurospora crassa* (Tyler 1987) and *Drosophila melanogaster* (Sharp and Garcia 1988), and the typical TTTT termination signal mentioned above would not be available for these retrotransposed copies, so the functionality of these 5S rRNA coding regions is, to us, unclear.

In CHAPTER 4.1.2, the linkage among 5S rDNA and other non-coding RNAs was studied for the first time in 97 metazoan genomes. No linkage was found to be stable over long evolutionary time in the Metazoa, but several non-coding RNA gene families were found to be tightly linked (within a distance of 10000 base pairs or less) to 5S rDNA in several taxa. Despite the new data provided in this thesis on the genomic organisation of 5S rDNA, why 5S rDNA often appears linked to other multigene families remains open. The high number of reported linkages of 5S rDNA to other multigene families (Drouin and Moniz de Sá 1995; CHAPTERS 4.1.1 and 4.1.2) can be a consequence of (1) the 'overstudy' of 5S rDNA compared to other non-coding RNA gene families; (2) the small size of some repeats, that enable the amplification and sequencing of 5S rDNA clones containing other multigene families in between (as in CHAPTER 4.1.1); (3) its high copy number, that makes it more likely to get transposed through unequal crossing-over; and (4) its high transcription levels, that makes it more likely to get transposed through an RNA intermediate. In fact, the possible higher *ability* of 5S rDNA to get transposed and linked to other multigene families could be addressed by studying the linkage among other non-coding RNA gene families using the methods described in CHAPTER 4.1.2 and Marz and Höner zu Siederdissen (in preparation), where the importance of copy number and transcription levels could also be accounted for.

6.2 Contributions of this thesis to the management of *Ensis* populations

The colonisation of European coastal waters by *E. directus* has been very successful. 35 years after it was found for the first time in the German Bight, this alien species is now present along most of the European Atlantic coast, including some isolated areas of Great Britain, and a population off the Asturian coast (north Iberia). Interestingly, a recent study concluded that this species integrated well into the existing community without suppressing native species (Dannheim and Rumohr 2012).

The higher genetic variation that we found in the introduced populations of *E. directus*, compared to that observed in two native populations (CHAPTER 4.3) can be related to the high colonisation capacity of the species. However, it should be noted that neutral genetic variation is of limited importance in assessing the ability to respond to new environments because it can underestimate non-neutral variation, and demographic bottlenecks might have positive effects on introduced

populations (e.g. by purging deleterious alleles) (Roman and Darling 2007). In fact, high levels of neutral genetic variation have been shown not to be essential for a sucessful introduction (Rollins et al. 2013).

In CHAPTER 4.3, in addition to studying genetic variation within and among *E. directus* populations in native and introduced ranges, we discovered a new *Ensis* species off Newfoundland (Canada) based on genetic and morphometric evidences. This study was complemented by the one in CHAPTER 4.4, in which we explored the species boundaries among Atlantic *Ensis* species by analysing nuclear and mitochondrial genomic regions.

de Queiroz (2007) proposed a unified species concept that defines species as separately evolving metapopulation lineages. Several recent studies on speciation, taxonomy, and systematics (e.g. Pfenninger et al. 2010; Marrone et al. 2010; Lega et al. 2012; Barata et al. 2012) have adopted this definition of species, that highlights the common element found in previous species concepts. In the context of a unified species concept, any property that provides evidence of lineage separation is relevant to inferring the boundaries and numbers of species (de Queiroz 2007). Considering that the analysis of nucleotide variation at certain genomic regions among closely related taxa is a useful approach to support or reject lineage separation, this strategy has become very common in species delimitation studies. Since the outcome of the species delimitation analysis can vary depending on the genomic regions under study (results based on different regions need not be compatible), it is common practice nowadays to analyse various regions in order to obtain as much information as possible. In metazoans, at least one mitochondrial and one nuclear region are usually sequenced, and results are interpreted in the light of other data (e.g. morphology, behaviour, geography, ecology, cytogenetics), following the integrative taxonomic approach (Dayrat 2005).

In CHAPTER 4.4 we used both phylogenetic trees and networks, and the GMYC method (General mixed Yule coalescent model; Pons et al. 2006), to test whether extant Atlantic *Ensis* morphospecies were separately evolving lineages. When only a few species are involved, reconstructing phylogenetic trees on which species are then defined as terminal clades is a widely used species delimitation method (Coissac et al. 2012). In turn, the GMYC method has also been widely used during the last few years. It tests for a significant shift in the branching rate in an ultrametric tree. Such a shift is indicative of the switch from between-species to within-species processes, expected if a sample comprises multiple individuals from a set of independently evolving species (Tang et al. 2012). It should be noted that the GMYC method may be prone to over delimitation (Carstens et al. 2013), and that both GMYC and phylogenetic tree reconstruction methods require that species have reached reciprocal monophyly.

Our results, based on the analysis of nuclear and mitochondrial regions in Atlantic *Ensis* (CHAPTER 4.4), were a rather easy case of species delimitation because morphospecies resulted to have reached reciprocal monophyly according to the majority of the reconstructed phylogenetic trees (see CHAPTER 4.4 for details), and reciprocal monophyly was further supported by the networks and by morphology. However, it is worth mentioning that the maximum likelihood method failed more often to recover reciprocal monophyly, compared to the Bayesian method.

Networks are used to represent evolutionary relationships among biological entities when these relationships are not tree-like (usually due to processes such as hybridisation, introgression, etc). In a splits graph, each edge represents a bipartition of the data, based on one or more characteristics; when bipartitions conflict, then, the algorithm creates a reticulation. Reticulations are parallelograms, with opposite edges representing the same split. The length of the edges represents split support (Morrison 2011). The neighbor-net networks reconstructed in CHAPTER 4.4 showed a rather 'tree-like' structure, meaning that there was no strong character conflict within the same datasets.

The discovery of *E. terranovensis* (CHAPTER 4.3), that is a closely related species to *E. directus* (CHAPTER 4.4), and the availability of the *E. directus* karyotype (CHAPTER 4.2) makes the karyotyping of the Canadian species particularly interesting. The comparison of both karyotypes would represent supplementary evidence on the degree of divergence between these razor shell species, and should be considered as a possibility for further research.

Some of the results of this thesis have practical consequences for the management of *Ensis* populations in various geographic regions. The discovery of *E. terranovensis* in Newfoundland (Canada) (CHAPTER 4.3) has pleasantly surprised marine biologists from that area, who had never noticed its differences with its sister species *E. directus*. Indeed, research is still to be done in order to unveil the distributional limits of both taxa, and whether or not they co-occur. Nonetheless, our results are relevant for the management plans of these species in the north eastern coast of the United States and in eastern Canada, where there is an increasing interest on *Ensis* fisheries motivated by a recent increase in the market value for razor shells, that has also resulted in increased interest in their culture (Flanagan 2013).

Similarly, the quite high mitochondrial divergence found among individuals of the species *E. macha* from an Atlantic and a Pacific population (CHAPTER 4.4) points towards the necessity of studying the populations of this intensively fished species along the coasts of Peru, Chile, and Argentina, in order to define evolutionarily significant units.

Finally, the confirmation of *E. siliqua* and *E. minor* as separate species (CHAPTERS 4.2 and 4.4) has practical consequences for the management of *Ensis* populations in various European areas. In Galicia, the confirmation of the co-occurrence of both species, together with *E. magnus*, makes it necessary to determine which species are present in the razor shell beds, in order to properly manage their populations, and to enable the traceability of this shell fish along the production chain. For this reason, some of the results of this thesis have been made available to the Galician Ministry for the Sea and Rural Environment.

As it can be noted from CHAPTER 5 (CORRIGENDUM), where we recorded some identification errors observed after the publication of some of our research articles, it is quite important to study species boundaries prior to study other biological features of the target organismal group. In fact, in our case, CHAPTER 4.4 should had been the first one to be published, but that was not possible because the majority of the specimens studied became available during the second half of the thesis period. Fortunately, now that at least the Atlantic species are better characterised, ongoing studies on the life cycles and reproduction of commercial *Ensis* can be better understood and will definitely promote a better management of *Ensis* populations.

7 CONCLUSIONS

7. CONCLUSIONS

Main conclusions of CHAPTER 4.1: **Evolutionary studies of 5S ribosomal DNA**.

- 5S ribosomal DNA copies are linked to U1 small nuclear DNA copies in at least 10 Pharidae species, from four different genera, showing the same gene orientation. Whereas tandemly-arranged 5S ribosomal DNA copies flanked by U1 small nuclear DNA are present in at least two razor shell species, tandemly-arranged U1 small nuclear DNA repeats are probably not present in the razor shell genomes. The linkage between 5S ribosomal DNA and U1 small nuclear DNA was found in some animal genomes other than razor shells, but it is not widespread in metazoans.
- The NTS region of razor shells is highly polymorphic and it is characterised by high levels of intragenomic variation. Some NTS copies are more closely related to NTSs from other species (and genera) than to NTSs from the species they were retrieved from, suggesting birth-and-death evolution and ancestral polymorphism.
- Purifying selection seems to have played an important role in the maintenance of the RNA coding regions, as well as the conserved upstream and downstream regions. However, several metazoan species have different types of 5S ribosomal DNA coding regions in the same genome.
- The organisation of 5S ribosomal DNA in metazoans is very flexible because it can be organised (1) in clusters, linked to other non coding RNAs, (2) in homogeneous clusters, with similar NTS sequences, (3) in heterogeneous clusters, with divergent NTS sequences, (4) in clusters in which coding regions displayed opposite orientations and (5) as dispersed copies. Several species display more than one of these features.
- Birth-and-death processes, selection, homogenising mechanisms typically involved in concerted evolution, and horizontal gene transfer events seem to be responsible for the diversity of 5S ribosomal DNA in metazoans.

Main conclusions of CHAPTER 4.2: **Cytogenetic characterisation of the razor shells *Ensis directus* (Conrad, 1843) and *E. minor* (Chenu, 1843) (Mollusca: Bivalvia).**

- The three European *Ensis* studied so far in terms of cytogenetics show more similarities among them than compared to the American species *E. directus* (Conrad, 1843).

Nonetheless, clear karyotype differences exist between the morphologically similar European species *Ensis minor* (Chenu, 1843) and *E. siliqua* (Linné, 1758). All these species have a diploid chromosome number of 38.

– According to fluorescent *in situ* hybridisation, the major ribosomal genes are located on one submetacentric chromosome pair in both species. A weaker fluorescent signal (or no signal at all) of the 5S ribosomal DNA probe compared to that obtained using the major ribosomal genes probe supports a more dispersed organisation of 5S ribosomal DNA in these species.

Main conclusions of CHAPTER 4.3: **Population genetic analysis of *Ensis directus* unveils high genetic variation in the introduced range and reveals a new species from the NW Atlantic.**

– The neutral genetic variation of *Ensis directus* (Conrad, 1843) in the introduced range is higher than in the native populations studied.

– Genetic and morphometric analyses showed that the razor shells from the population of Long Pond (Newfoundland, Canada) belong to a new species that was named *E. terranovensis* Vierna & Martínez-Lage, 2012.

Main conclusions of CHAPTER 4.4: **Species delimitation and DNA barcoding of Atlantic *Ensis* (Bivalvia, Pharidae).**

– All extant Atlantic *Ensis* morphospecies are separately evolving lineages according to nuclear and mitochondrial DNA, being *Ensis* at each side of the Atlantic reciprocally monophyletic.

– The morphologically similar species *E. minor* (Chenu, 1843) and *E. siliqua* (Linné, 1758) co-occur along the NW Iberian coast. Thus, the commercial *Ensis* species which are found in Galicia are *E. minor*, *E. siliqua*, and *E. magnus* Schumacher, 1817. The non-commercial species *E. ensis* (Linné, 1758) also occurs in Galician waters.

– The analysis of a fragment of the mitochondrial cytochrome oxidase subunit I gene is capable of identifying *Ensis* specimens to the species level. This methodology, known as DNA barcoding, can be used for reliable and inexpensive identifications of specimens.

8 REFERENCES

8. REFERENCES

Alaska Department of Fish and Game (2010) Accessed at http://www.adfg.alaska.gov/index.cfm?adfg=razorclam.main on April 24th 2012.

Arias A, Fernández-Moreno M, Fernández-Tajes J, Gaspar MB, Méndez J (2011) Strong genetic differentiation among east Atlantic populations of the sword razor shell (*Ensis siliqua*) assessed with mtDNA and RAPD markers. Helgoland Marine Research 65:81-89.

Arias-Pérez A, Fernández-Tajes J, Gaspar MB, Méndez J (2012) Isolation of microsatellite markers and analysis of genetic diversity among east atlantic populations of the sword razor shell *Ensis siliqua*: a tool for population management. Biochemical Genetics 50:397-415.

Ariz Abarca L, Cortés Segovia C, González Yáñez J, Barahona Toledo N, Nilo Gatica M (2007) Situación actual de la pesquería del recurso huepo (*Ensis macha*) en la VII Región. Instituto de Fomento Pesquero (accessed at http://www.fip.cl/FIP/Archivos/pdf/informes/inffinal%202006-44.pdf).

Armonies W (1996) Changes in distribution patterns of 0-group bivalves in the Wadden Sea: byssus-drifting releases juveniles from constraints of hydrography. Journal of Sea Research 35:2323-2334.

Armonies W, Reise K (1999) On the population development of the introduced razor clam *Ensis americanus* near the island of Sylt (North Sea). Helgoländer Meeresuntersuchungen 52:291-300.

Arnheim N, Krystal M, Schmickel R, Wilson G, Ryder O, Zimmer E (1980) Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proceedings of the National Academy of Sciences 12:7323-7327.

Avellanal MH, Jaramillo E, Clasing E, Quijon P, Contreras H (2002) Reproductive cycle of the bivalves *Ensis macha* (Molina, 1782) (Solenidae), *Tagelus dombeii* (Lamarck, 1818) (Solecurtidae), and *Mulinia edulis* (King, 1831) (Mactridae) in southern Chile. The Veliger 45:33-44.

Barata M, Carranza S, Harris D (2012) Extreme genetic diversity in the lizard *Atlantolacerta andreanskyi* (Werner, 1929): A montane cryptic species complex. BMC Evolutionary Biology 12:167.

Barón PJ, Real LE, Ciocco NF, Ré ME (2004) Morphometry, growth and reproduction of an

Atlantic population of the razor clam *Ensis macha* (Molina, 1782). Scientia Marina 68:211-217.

Beauparlant MA, Drouin G (2013) Multiple independent insertions of 5S rRNA genes in the spliced-leader gene family of trypanosome species. Current Genetics 0:000-000. DOI: 10.1007/s00294-013-0404-z.

Bellavia D, Dimarco E, Naselli F, Caradonna F (2013) DNA-methylation dependent regulation of embryo-specific 5S ribosomal DNA cluster transcription in adult tissues of sea urchin *Paracentrotus lividus*. Genomics 102:397-402.

Beukema JJ, Dekker R (1995) Dynamics and growth of a recent invader into European coastal waters: the American razor clam, *Ensis directus*. Journal of the Marine Biological Association of the United Kingdom 75:351-362.

Caradonna F, Bellavia D, Clemente AM, Sisino G, Barbieri R (2007) Chromosomal localization and molecular characterization of three different 5S ribosomal DNA clusters in the sea urchin *Paracentrotus lividus*. Genome 50:867-870.

Cardoso JFMJ, Witte JIJ, van der Veer HW (2009) Reproductive investment of the American razor clam *Ensis americanus* in the Dutch Wadden Sea. Journal of Sea Research 62:295-298.

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. Molecular Ecology 22:4369-4383.

Coan EV, Valentich-Scott P (2012) Bivalve seashells of Tropical West America: Marine bivalve mollusks from Baja California to Northern Peru. Santa Barbara Museum of Natural History, California, USA.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Molecular Ecology 21:1834-1847.

Cosel von R (1990) An introduction to the razor shells (Bivalvia: Solenacea). In Morton (ed), The Bivalvia - Proceedings of a Memorial Symposium in Honour of Sir Charles Maurice Yonge, Edinburgh, 1986, pp: 283-305. Hong Kong University Press. Hong Kong.

Cosel von R (1993) The razor shells of the eastern Atlantic. Part 1: Solenidae and Pharidae I (Bivalvia: Solenacea). Archiv für Molluskenkunde 122:207-321.

Cosel von R (2009) The razor shells of the eastern Atlantic, part 2. Pharidae II: The genus *Ensis* Schumacher, 1817 (Bivalvia, Solenoidea). Basteria 73:1-48.

Da Costa F, Darriba S, Martínez-Patiño D (2008) Embryonic and larval development of *Ensis arcuatus* (Jeffreys, 1865) (Bivalvia: Pharidae). Journal of Molluscan Studies 74:103-109.

Dannheim J, Rumohr H (2012) The fate of an immigrant: *Ensis directus* in the eastern German Bight. Helgoland Marine Research 66:307-317.

Darriba S, San Juan F, Guerra A (2004) Reproductive cycle of the razor clam *Ensis arcuatus* (Jeffreys, 1865) in northwest Spain and its relation to environmental conditions. Journal of Experimental Marine Biology and Ecology 311:101-115.

Darriba S, San Juan F, Guerra A (2005) Gametogenic cycle of *Ensis siliqua* (Linnaeus, 1758) in the Ría de Corcubión, northwestern Spain. Journal of Molluscan Studies 71:47-51.

Dayrat B (2005) Towards integrative taxonomy. Biological Journal of the Linnean Society 85:407-415.

de Queiroz (2007) Species concepts and species delimitation. Systematic Biology 56: 879-886.

Del Piero D, Dacaprile R (1998) The alternating recruitment pattern in *Ensis minor*, an exploited bivalve in the Gulf of Trieste, Italy. Hydrobiologia 375-376:67-72.

Dimarco E, Cascone E, Bellavia D, Caradonna F (2012) Functional variants of 5S rRNA in the ribosomes of common sea urchin *Paracentrotus lividus*. Gene 508:21-25.

Drouin G, Moniz de Sá M (1995) The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. Molecular Biology and Evolution 12:481-493.

Eickbush TH, Eickbush DG (2007) Finely orchestrated movements: Evolution of the ribosomal RNA genes. Genetics 175: 477-485.

Eirín-López JM, Rebordinos L, Rooney AP, Rozas J (2012) The Birth-and-death evolution of multigene families revisited. In Garrido-Ramos MA (ed), Repetitive DNA. Genome Dynamics 7:170-196.

Espiñeira M, González-Lavín N, Vieites JM, Santaclara FJ (2009) Development of a method for the genetic identification of commercial bivalve species based on mitochondrial 18S rRNA sequences. Journal of Agricultural and Food Chemistry 57:495-502.

Espinoza R, Tarazona J, Laudien J (2010) Overfishing population characteristics of razor clam, *Ensis macha*, from Independencia Bay, Peru, in 2004 year. Revista Peruana de Biología 17:285-292.

Essink K (1985) On the occurrence of the American jack-knife clam *Ensis directus* (Conrad, 1843) (Bivalvia, Cultellidae) in the Dutch Wadden Sea. Basteria 49:73-80.

Fahy E, Gaffney J (2001) Growth statistics of an exploited razor clam (*Ensis siliqua*) bed at Gormanstown, Co Meath, Ireland. Hydrobiologia 465:139-151.

Fernández-Tajes J, Freire R, Méndez J (2010) A simple one-step PCR method for the identification between European and American razor clams species. Food Chemistry 118:995-998.

Fernández-Tajes J, Gaspar MB, Méndez J (2012) identification of *Ensis siliqua* samples and establishment of the catch area using a species-specific microsatellite marker. Journal of AOAC International 95:820-823.

Fernández-Tajes J, González-Tizón A, Martínez-Lage A, Méndez J (2003) Cytogenetics of the razor clam *Solen marginatus* (Mollusca: Bivalvia: Solenidae). Cytogenetic and Genome Research 101:43-46.

Fernández-Tajes J, Martínez-Lage A, Freire R, Guerra A, Méndez J, González-Tizón AM (2008) Genome sizes and karyotypes in the razor clams *Ensis arcuatus* (Jeffreys, 1865) and *E. siliqua* (Linnaeus, 1758). Cahiers de Biologie Marine 49:79-85.

Fernández-Tajes J, Méndez J (2007) Identification of the razor clam species *Ensis arcuatus*, *E. siliqua*, *E. directus*, *E. macha*, and *Solen marginatus* using PCR-RFLP analysis of the 5S rDNA region. Journal of Agricultural and Food Chemistry 55:7278-7282.

Fernández-Tajes J, Méndez J (2009) Two different size classes of 5S rDNA units coexisting in the same tandem array in the razor clam *Ensis macha*: is this region suitable for phylogeographic studies? Biochemical Genetics 47:775-788.

Flanagan MP (2013) Investigation of early development and importance of sediment choice in the hatchery production of razor clams, *Ensis directus*. Honors College. Paper 111. http://digitalcommons.library.umaine.edu/honors/111

Francisco-Candeira M, González-Tizón A, Varela MA, Martínez-Lage A (2007) Development of microsatellite markers in the razor clam *Solen marginatus* (Bivalvia: Solenidae). Journal of the Marine Biological Association of the United Kingdom 87:977-978.

Freire R, Fernández-Tajes J, Méndez J (2008) Identification of razor clams *Ensis arcuatus* and *Ensis siliqua* by PCR–RFLP analysis of ITS1 region. Fisheries Science 74:511-515.

Fujiwara M, Inafuku J, Takeda A, Watanabe A, Fujiwara A, Kohno S, Kubota S (2009) Molecular

organization of 5S rDNA in bitterlings (Cyprinidae). Genetica 135:355-365.

Gaspar MB, Monteiro CC (1998) Reproductive cycles of the razor clam *Ensis siliqua* and the clam *Venus striatula* off Vilamoura, southern Portugal. Journal of the Marine Biological Association of the United Kingdom 78:1247-1258.

González-Romero R, Ausió J, Méndez J, Eirín-López JM (2009) Histone genes of the razor clam *Solen marginatus* unveil new aspects of linker histone evolution in protostomes. Genome 52:597-607.

Gornung E, Colangelo P, Annesi F (2007) 5S ribosomal RNA genes in six species of Mediterranean grey mullets: genomic organization and phylogenetic inference. Genome 50:787-795.

Guo X, Ford SE, Zhang F (1999) Molluscan aquaculture in China. Journal of Shellfish Research 18:19-31.

Hassan R, Kanakaraju D (2013) Razor clams (class Bivalvia) of Kuala Selangor, Malaysia: morphology, genetic diversity and heavy metal concentration. Borneo Journal of Resource Science and Technology 2:19-27.

Henderson SM, Richardson CA (1994) A comparison of the age, growth rate and burrowing behaviour of the razor clams, *Ensis siliqua* and *E. ensis*. Journal of the Marine Biological Association of the United Kingdom 74:939-954.

Hmida L, Fassatoui C, Ayed D, Ayache N, Romdhane MS (2012) Genetic characterization of the razor clam *Solen marginatus* (Mollusca: Bivalvia: Solenidae) in Tunisian coasts based on isozyme markers. Biochemical Systematics and Ecology 40:146-155.

Holme NA (1954) The ecology of British species of *Ensis*. Journal of the Marine Biological Association of the United Kingdom 33:145-172.

Huber M (2010) Compendium of bivalves. A full-color guide to 3,300 of the World's Marine Bivalves. A status on Bivalvia after 250 years of research. ConchBooks, Hackenheim, Germany.

Jiang Q, Li Q, Yuan Y, Kong LF (2010) Development and characterization of 14 polymorphic microsatellite loci in the razor clam (*Sinonovacula constricta*). Conservation Genetics Resources 2:81-83.

Krakau M, Thieltges DW, Reise K (2006) Native parasites adopt introduced bivalves of the North Sea. Biological Invasions 8:919-925.

Lebour MV (1938) Notes on the breeding of some lamellibranchs from Plymouth and their larvae. Journal of the Marine Biological Association of the United Kingdom 23:119-144 .

Lega M, Fior S, Prosser F, Bertolli A, Li M, Varotto C (2012) Application of the unified species concept reveals distinct lineages for disjunct endemics of the *Brassica repanda* (Brassicaceae) complex. Biological Journal of the Linnean Society 106:482-497.

Luczak C, Dewarumez J-M, Essink K (1993) First record of the American jack knife clam *Ensis directus* on the French coast of the North Sea. Journal of the Marine Biological Association of the United Kingdom 73:233-235.

Marrone F, Brutto SL, Arculeo M (2010) Molecular evidence for the presence of cryptic evolutionary lineages in the freshwater copepod genus *Hemidiaptomus* GO Sars, 1903 (Calanoida, Diaptomidae). Hydrobiologia 644:115-125.

Martínez D (2002) Estudio de los solénidos, *Solen marginatus* (Pennánt, 1777) y *Ensis siliqua* (Linné, 1758), de los bancos naturales de la Ría de Ortigueira y Ría del Barquero: ciclo gametogénico, composición bioquímica y cultivo larvario. PhD thesis, University of Santiago de Compostela, Spain.

Marz M, Höner zu Siederdissen C (In preparation). Statistical approach for evolutionary gene linkage detection.

Morrison DA (2011) An introduction to phylogenetic networks. RJR Productions. Uppsala, Sweden.

Nei M, Hughes AL (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In 11th Histocompatibility Workshop and Conference, K Tsuji, M Aizawa, T Sasazuki (eds), pp; 27-38. Oxford University Press. Oxford, United Kingdom.

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics 39:121-152.

Nelson DW, Linning RM, Davison PJ, Honda BM (1998) 5'-flanking sequences required for efficient transcription in vitro of 5S RNA genes, in the related nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Gene 218: 9-16.

Niu D, Wang L, Sun F, Liu Z, Li J (2013) Development of molecular resources for an intertidal clam, *Sinonovacula constricta*, using 454 transcriptome sequencing. PloS ONE 8:e67456.

Niu DH, Feng BB, Liu DB, Zhong YM, Shen HD, Li JL (2012) Significant genetic differentiation among ten populations of the razor clam *Sinonovacula constricta* along the coast of China

revealed by a microsatellite analysis. Zoological Studies 51:406-414.

Niu DH, Li JL, Feng BB, Liu DB (2009) ISSR analysis on genetic structure of six *Sinonovacula constricta* populations. Chinese Journal of Applied & Environmental Biology 3:011.

Niu DH, Li JL, Shen HD, Jiang ZY (2008) Sequence variability of mitochondrial DNA-COI gene fragment and population genetic structure of six *Sinonovacula constricta* populations. Acta Oceanologica Sinica 30:109-116.

Niu DH, Li JL, Liu DB (2008) Polymorphic microsatellite loci for population studies of the razor clam, *Sinonovacula constricta*. Conservation Genetics 9:1393-1394.

Niu DH, Li JL, Wang GL, Jiang ZY, Zhang WB, Shen YB, Feng BB (2007) The genetic diversity of mitochondrial 16S rRNA gene fragment in six populations of *Sinonovacula constricta.* Journal of Shanchai Fisheries University 1:1-6.

Perina A, Seoane D, González-Tizón A, Rodríguez-Fariña F, Martínez-Lage A (2011) Molecular organization and phylogenetic analysis of 5S rDNA in crustaceans of the genus *Pollicipes* reveal birth-and-death evolution and strong purifying selection. BMC Evolutionary Biology 11:304.

Peterson RC, Doering JL, Brown DD (1980) Characterization of two *Xenopus* somatic 5S DNAs and one minor oocyte-specific 5S DNA. Cell 20:131-141.

Pfenninger M, Véla E, Jesse R, Elejalde MA, Liberto F, Magnin F, Martínez-Ortí A (2010) Temporal speciation pattern in the western Mediterranean genus *Tudorella* P. Fischer, 1885 (Gastropoda, Pomatiidae) supports the Tyrrhenian vicariance hypothesis. Molecular Phylogenetics and Evolution 54:427-436.

Pinhal D, Yoshimura TS, Araki CS, Martins C (2011) The 5S rDNA family evolves through concerted and birth-and-death evolution in fish genomes: an example from freshwater stingrays. BMC Evolutionary Biology 11:151.

Plataforma Tecnolóxica da Pesca (2013) Accessed at http://pescadegalicia.com/ on July 1st 2013.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Systematic Biology 55:595-609.

Rollins LA, Moles AT, Lam S, Buitenwerf R, Buswell JM, Brandenburger CR, Flores-Moreno H, Nielsen KB, Couchman E, Brown GS, Thomson FJ, Hemmings F, Frankham R, Sherwin

WB (2013) High genetic diversity is not essential for successful introduction. Ecology and Evolution 0:000-000. DOI: 10.1002/ece3.824.

Roman J, Darling JA (2007) Paradox lost: genetic diversity and the success of aquatic invasions. Trends in Ecology & Evolution 22:454-464.

Rooney AP (2004) Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans. Molecular Biology and Evolution 21:1704-1711.

Rooney AP, Ward TJ (2005) Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proceedings of the National Academy of Sciences of the United States of America 102:5084-5089.

Sharp SJ, Garcia AD (1988) Transcription of the *Drosophila melanogaster* 5S RNA gene requires an upstream promoter and four intragenic sequence elements. Molecular and Cellular Biology 8:1266-1274

Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences 109:16208-16212.

Tyler BM (1987) Transcription of *Neurospora crassa* 5 S rRNA genes requires a TATA box and three internal elements. Journal of Molecular Biology 196:801-811.

Van Urk RM (1964) The genus *Ensis* in Europe. Basteria 28:13-44.

Van Urk RM (1987) *Ensis americanus* (Binney) (syn. *E. directus* auct. non Conrad). A recent introduction from Atlantic North-America. Journal of Conchology 32:329-333.

Varela MA, González-Tizón A, Francisco-Candeira M, Martínez-Lage A (2007) Isolation and characterization of polymorphic microsatellite loci in the razor clam *Ensis siliqua*. Molecular Ecology Notes 7:221-222.

Varela MA, Martínez-Lage A, González-Tizón AM (2009) Temporal genetic variation of microsatellite markers in the razor clam *Ensis arcuatus* (Bivalvia: Pharidae). Journal of the Marine Biological Association of the United Kingdom 89:1703-1707.

Varela MA, Martínez-Lage A, González-Tizón AM (2012) Genetic heterogeneity in natural beds of the razor clam *Ensis siliqua* revealed by microsatellites. Journal of the Marine Biological

Association of the United Kingdom 92:171-177.

Vierna J, González-Tizón AM, Martínez-Lage A (2009) Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). Biochemical Genetics 47:635-644.

Vizoso M, Vierna J, González-Tizón AM, Martínez-Lage A (2011) The 5S rDNA gene family in mollusks: characterization of transcriptional regulatory regions, prediction of secondary structures, and long-term evolution, with special attention to Mytilidae mussels. Journal of Heredity 102:433-447.

Wang J, Li Q, Kong L (2010) Genetic variation and differentiation in wide ranging populations of razor clam (*Sinonovacula constricta*) inferred from AFLP markers. Journal of Ocean University of China 9:297-302.

Wang J, Zhao X, Zhou L, Xiang J (1998) Chromosome study of *Sinonovacula constricta* (Bivalvia). Oceanologia et Limnologia Sinica 29:191-196.

Yuan Y, Li Q, Kong L, Yu H (2012) The complete mitochondrial genome of *Solen strictus* (Bivalvia: Solenidae). Mitochondrial DNA 23:112-114.

Yuan Y, Li Q, Kong L, Yu H (2012) The complete mitochondrial genome of the grand jackknife clam, *Solen grandis* (Bivalvia: Solenidae): a novel gene order and unusual non-coding region. Molecular Biology Reports 39:1287-1292.

Yuan Y, Li Q, Yu H, Kong L (2012) The complete mitochondrial genomes of six heterodont bivalves (Tellinoidea and Solenoidea): variable gene arrangements and phylogenetic implications. PloS ONE 7: e32353.

Zeng GQ, Fang J, Jia SJ, Zhang YP, Chen C, Zheng YY, Yu JQ (2010) Biochemical genetic analysis of eight isozymes in intra-populations of razor clam *Cultellus attenuatus*. Fisheries Science 11:011.

9 CURRICULUM VITAE

9. CURRICULUM VITAE

Name: Joaquín Vierna Fernández. **Place of birth**: Ortigueira, La Coruña, Spain (March, 1983).

Bachelor in Biology ('licenciatura'), Universidade da Coruña 2001-2006. **Erasmus** exchange student in Università degli Studi di Padova (Italy) 2004-2005. **Master** in Genetics ('DEA'), Universidade da Coruña 2006-2008.

Spanish / **Galician** mother tongue. **English** C2 level. **Italian** C2 level.

Current positions: PhD student at Evolutionary Biology Group, Universidade da Coruña (since September 2006) and general manager at AllGenetics (since October 2011).

Address: Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, A Fraga 10, E-15008 La Coruña, Spain. **E-mail**: jvierna@udc.es.

Publications

Joaquín Vierna, Joël Cuperus, Andrés Martínez-Lage, Jeroen M. Jansen, Alejandra Perina, Hilde Van Pelt, Ana M. González-Tizón (2013). Species delimitation and DNA barcoding of Atlantic *Ensis* (Bivalvia, Pharidae). Zoologica Scripta doi:10.1111/zsc.12038.

Joaquín Vierna, Stefanie Wehner, Andrés Martínez-Lage, Manja Marz (2013) Systematic analysis and evolution of 5S ribosomal DNA in metazoans. Heredity 111:410-421.

Ana M. González-Tizón, Verónica Rojo, **Joaquín Vierna**, K. Thomas Jensen, Emilie Egea, Andrés Martínez-Lage (2013) Cytogenetic characterisation of the razor shells *Ensis directus* (Conrad, 1843) and *E. minor* (Chenu, 1843) (Mollusca: Bivalvia). Helgoland Marine Research 67:73-82.

Joaquín Vierna, K. Thomas Jensen, Ana M. González-Tizón, Andrés Martínez-Lage (2012) Population genetic analysis of *Ensis directus* unveils high levels of genetic variation in the introduced range, and reveals a new species from the NW Atlantic. Marine Biology 159:2209-2227.

Miguel Vizoso, **Joaquín Vierna,** Ana M. González-Tizón, Andrés Martínez-Lage (2011) The 5S rDNA gene family in mollusks: characterization of transcriptional regulatory regions, prediction of secondary structures, and long-term evolution, with special attention to Mytilidae mussels. Journal of Heredity 102:433-447.

Joaquín Vierna, K. Thomas Jensen, Andrés Martínez-Lage, Ana. M. González-Tizón (2011) The linked units of 5S rDNA and U1 snDNA of razor shells (Mollusca: Bivalvia: Pharidae). Heredity 107:127-142.

Joaquín Vierna, Andrés Martínez-Lage, Ana M. González-Tizón (2010) Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. Genome 53:23-34.

Joaquín Vierna, Ana M. González-Tizón, Andrés Martínez-Lage (2009) Long-Term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). Biochemical Genetics 47:635-644.

Grants

2010: Travel grant from the Universidade da Coruña to perform a 6-week internship at the RNA Bioinformatics Group, Department of Pharmaceutical Chemistry, Philipps-Universität Marburg (Germany).

2009 - 2011: 'María Barbeito' fellowship for pre-doctoral researchers (Galician Regional Government).

2009: Travel grant from the Galician Regional Government to perform an 8-week internship at the Institute of Biological Sciences, Aarhus Universitet (Denmark).

2006 - 2008: DEA students' fellowship ('bolsa de terceiro ciclo', Galician Regional Government).

2006: Collaboration grant for last-year students at the Department of Molecular and Cell Biology (Genetics), Faculty of Science, Universidade da Coruña (Spain), funded by the Spanish Government.

Research internships

11-12/2010 RNA Bioinformatics Group, Department of Pharmaceutical Chemistry, Philipps-Universität Marburg (Germany). Six weeks.

11-12/2009 Institute of Biological Sciences. Aarhus Universitet, Aarhus (Denmark). Eight weeks.

07/2009 Centre d'Océanologie de Marseille, Marseilles (France). Two weeks.

06/2008 Institute of Biological Sciences. Aarhus Universitet, Aarhus (Denmark). One week.

11-12/2007 Department of Systematics and Evolution. Muséum national d'Histoire naturelle, Paris (France). Two weeks.

Courses given

May 2011 and May 2010 - Biology degree (5th year): Practical course on Human Genetics (40h). DNA extraction; PCR; gel electrophoresis; haplotype networks.

October 2010 - Biology degree (5th year): Practical course on Genetic Methods (50h). GenBank/EMBL/DDBJ databases; sequence alignment; polymorphism and divergence;

phylogenetic trees; phylogenetic networks.

October 2010 - Biology degree (3rd year): Practical course on Genetics (12h). *Drosophila* lesson: Lab culture procedures; manipulating the animals in the lab; differentiating males, females, and mutants.

Most relevant courses attended

09/2010 Marine Macrofauna – systematics and taxonomy. Sven Lovén Centre for Marine Sciences (Tjärnö, Sweden). 9 days.

03/2009 Bodega Applied Phylogenetics Workshop. Organised by UC Davies (California, USA) at the Bodega Marine Lab. 7 days.

07/2007 Analyzing Biodiversity and Life History Strategies. Organised by Marine Genomics Europe at the Sven Lovén Centre for Marine Sciences (Kristineberg, Sweden). 5 ECTS.

Popularisation of science activities

03/06/2010 Invited talk at 'Casa de las Ciencias' (La Coruña Science Museum) during its XXV anniversary celebrations. 'El difícil caso de las navajas que se parecían demasiado...' (The difficult case of the razor shells that were too much alike...).

05/2010 Organiser of several activities during the 'XV Día de la ciencia en la calle' (XV Science in the street day) together with other colleagues from Precarios-Galicia, the association of Galician young researchers. La Coruña (Spain).

Book chapter: 'Pingelap, la vida en blanco y negro' (Pingelap, a black and white life) in 'Aquí estamos... mutando'. Pp 169-176. Ed. Ana M. González Tizón (2009) ISBN 978-84-9749-330-7.

Radio program: co-founder of Do The Evolution, the biology and research radio show at CUAC FM station.

Organisation of scientific events

03/2010 II School of Scientific Research - 'II Taller de Investigación Científica: Una profesión ¿con futuro?' (Faculty of Sciences, Universidade da Coruña). Member of organising committee.

09/2009 II Meeting of the Galician Network for Conservation of Biological Diversity (Faculty of Sciences, Universidade da Coruña). Member of organising committee.

04/2009 I School of Scientific Research - 'I Taller de Investigación Científica: Cómo investigar sin morir en el intento' (Faculty of Sciences, Universidade da Coruña). Member of organising

committee.

Membership of scientific societies

Sociedad Española de Biología Evolutiva (SESBE), the Spanish Society for Evolutionary Biology.

Precarios-Galicia, the association of Galician young researchers (vice-chairman April 2009-June 2011).

Federación de Jóvenes Investigadores (FJI/Precarios), the federation of Spanish young researchers

10 APPENDIX

10. APPENDIX

10.1 Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia)

Joaquín Vierna, Ana M. González-Tizón, Andrés Martínez-Lage (2009) Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). Biochemical Genetics 47:635–644.

Bibliometrics 2012 JCR Science Edition

Impact factor: 0.938

Biochemistry & Molecular Biology: Q4

Genetics & Heredity: Q4

Biochem Genet (2009) 47:635–644
DOI 10.1007/s10528-009-9255-1

Long-Term Evolution of 5S Ribosomal DNA Seems to Be Driven by Birth-and-Death Processes and Selection in *Ensis* Razor Shells (Mollusca: Bivalvia)

Joaquín Vierna · Ana M. González-Tizón · Andrés Martínez-Lage

Received: 12 October 2008 / Accepted: 6 February 2009 / Published online: 25 July 2009

Abstract A study of nucleotide sequence variation of 5S ribosomal DNA from six *Ensis* species revealed that several 5S ribosomal DNA variants, based on differences in their nontranscribed spacers (NTS), occur in *Ensis* genomes. The 5S rRNA gene was not very polymorphic, compared with the NTS region. The phylogenetic analyses performed showed a between-species clustering of 5S ribosomal DNA variants. Sequence divergence levels between variants were very large, revealing a lack of sequence homogenization. These results strongly suggest that the long-term evolution of *Ensis* 5S ribosomal DNA is driven by birth-and-death processes and selection.

Keywords 5S ribosomal DNA · Birth-and-death evolution · *Ensis* · Mollusca · Bivalvia

Introduction

Nuclear ribosomal DNA (nrDNA) is composed of the major ribosomal genes, their spacers, the 5S ribosomal gene, and its spacer. These genomic regions has been widely used in molecular evolutionary studies, taken as an object of study itself, in an attempt to understand the evolutionary forces that shaped extant variation, or used as molecular markers in phylogenetic surveys at different levels.

J. Vierna · A. M. González-Tizón · A. Martínez-Lage (✉)
Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, A Zapateira s/n, La Coruña 15071, Spain
e-mail: andres@udc.es

J. Vierna
e-mail: jvierna@udc.es

A. M. González-Tizón
e-mail: hakuna@udc.es

The major ribosomal genes and spacers are tandemly repeated in the genome, forming long arrays, and each eukaryote species may have one or more of these arrays.

The 5S ribosomal DNA (5S rDNA) is formed by the 5S rRNA gene and a nontranscribed spacer (NTS). It is organized in the same tandemly repeated units as the major ribosomal genes in many species (Rooney 2004; Rooney and Ward 2005). However, it can also be found dispersed throughout the genome (in *Schizosaccharomyces pombe*, Wood et al. 2002, and in other fungi belonging to the subphylum Pezizomycotina, Rooney and Ward 2005), in tandem arrays, separate from the major ribosomal genes, or in both types of arrangements (in *Homo sapiens*, Little and Braaten 1989). In relation to copy number, each species genome contains many repeats of 5S rDNA (e.g., 9,000–24,000 in *Xenopus*, Brown and Sugimoto 1974; 600–4,600 in soybean, Danna et al. 1996; ~686 in rice, Li et al. 2002).

The evolution of nrDNA is usually explained by the so-called model of concerted evolution: mutations occurring in the ribosomal genes or in their spacers can spread to the other units of the array through the action of homologous and nonhomologous unequal crossovers and gene conversions. The molecular mechanism of gene conversion in multigene families is not well understood (Nei and Rooney 2005), but there is a commonly held opinion that a combination of both unequal crossovers and gene conversions gives rise to the concerted evolution of ribosomal DNA in all organisms (Eickbush and Eickbush 2007). Unequal crossovers and gene conversions have a homogenizing effect, and they significantly reduce the sequence divergence between members of the same array or between members of different arrays. Other evolutionary forces that reduce the variation produced by mutations in nrDNA are genetic drift and selection (selection is expected to have a much greater effect over the genes than over the spacers).

In several animal phyla, however, examples have been reported in which sequence divergence levels between ribosomal genes or spacers seem to be much higher than expected under a strict concerted evolution scenario (e.g., Insua et al. 2001; Leo and Barker 2002; Freire et al. 2005; Keller et al. 2006; Sword et al. 2007; Caradonna et al. 2007; López-Piñón et al. 2008).

Previous studies have demonstrated that nrDNA may evolve under the birth-and-death model (Nei and Hughes 1992). Under this model, new ribosomal gene copies are created by gene duplication, during the evolution of a particular group of organisms. These new copies can persist in the genomes as functional genes for a long time, or become pseudogenes and accumulate deleterious mutations. Organisms whose nrDNA (e.g., the 5S rDNA) is evolving under birth-and-death processes will accumulate several variants of (in this case) 5S ribosomal genes and spacers.

There are examples in which purifying or positive selection and birth-and-death processes are responsible for the evolution of 5S rDNA (Rooney and Ward 2005; Fujiwara et al. 2009). The 5S rDNA of these organisms is characterized by the occurrence of several 5S rDNA variants in their genomes, with low levels of nucleotide variation in the 5S rRNA gene (maintained by selection) and a highly polymorphic NTS region.

Under a birth-and-death scenario, 5S rDNA variants sampled from different species of a given group are expected to form clusters based on their sequence similarities in a phylogenetics tree. These sequences will not cluster by species (as

would be expected under concerted evolution). Rather, they form several clades composed of different species (see Fig. 1 in Rooney and Ward 2005).

According to Rooney and Ward (2005), birth-and-death evolution can be detected by two means: a phylogenetic analysis of multigene family members, which will reveal a between-species gene clustering pattern; and an examination of gene sequence divergence levels, in which case a relatively high proportion of changes between gene family members will indicate a lack of homogenization. In the present study we performed such analyses in order to understand the long-term evolution of 5S rDNA in *Ensis* razor shells (Mollusca: Bivalvia). The high levels of 5S rDNA sequence divergence that we found in the species *E. directus* led us to analyze 5S rDNA variation in a broader phylogenetic context. We used sequences from five other *Ensis* species and found evidence of birth-and-death evolution of 5S rDNA, for the first time in molluscs.

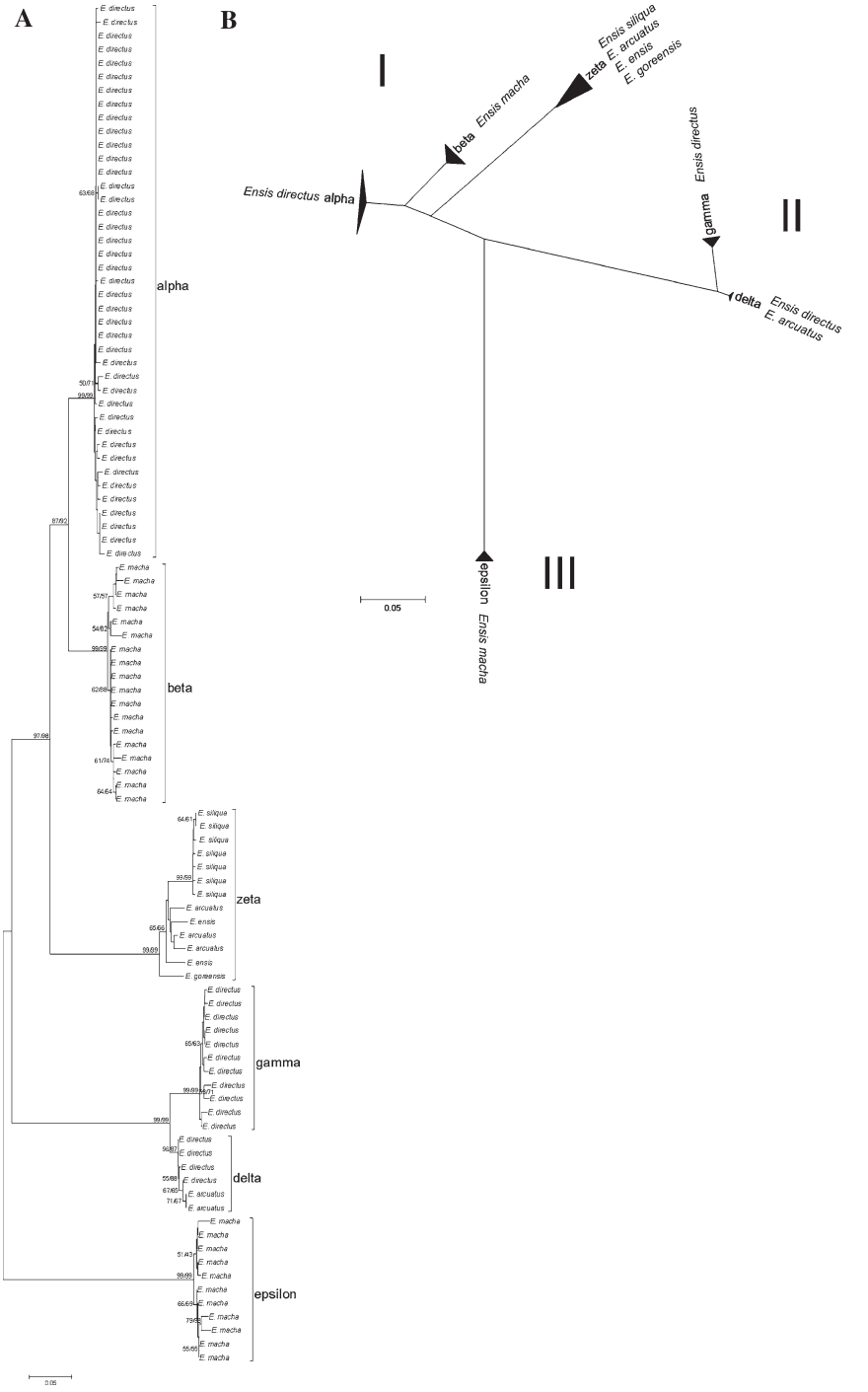
Materials and Methods

We analyzed six species of *Ensis* (Schumacher 1817) from the Atlantic coast and Chile: *E. directus* (Conrad 1843), *E. macha* (Molina 1792), *E. arcuatus* (Jeffreys 1865), *E. ensis* (Linné 1758), *E. siliqua* (Linné 1758), and *E. goreensis* (Clessin 1888). Specimens were preserved in ethanol, except in the case of *E. goreensis*, for which we used dried museum material (specimen code MNHN17948).

Extraction of genomic DNA was performed from foot tissue using the NucleoSpin Tissue kit (Macherey–Nagel and Co.). For PCR amplifications a pair of primers was designed from *Mytilus edulis* and *M. galloprovincialis* 5S rRNA gene sequences available in the international nucleotide sequence databases (DDBJ/EMBL/GenBank, accession nos. AJ312081–AJ312087; AJ312075–AJ312080) using GeneFisher (Giegerich et al. 1996): 5S-Univ-F (5′-ACCGGTGTTTTCAACGTGAT) and 5S-Univ-R (5′-CGTCCGATCACCGAAGTTAA). Primers had opposite orientation and were separated by 3 bp. Each reaction was performed in a total volume of 25 μl containing ~25 ng genomic DNA, 0.625 U *Taq* DNA polymerase (Roche Diagnostics), 5 nmol each dNTP (Roche Diagnostics), 20 pmol each primer, and the buffer recommended by the polymerase supplier. PCR conditions were as follows: an initial denaturation step at 94°C for 3 min, followed by 35 cycles of denaturation at 94°C for 45 s, annealing at 50°C for 45 s, extension at 72°C for 2 min, and final extension at 72°C for 10 min. PCR products were run on agarose gels, stained with ethidium bromide, and imaged under UV fluorescence. Bands ranging from ~450 bp to ~1050 bp were excised from the gel and purified using a DNA Gel Extraction Kit (Millipore). Bands were cloned using Topo TA Cloning (Invitrogen). A QiaPrep Spin Miniprep Kit (Qiagen) was used to purify the plasmids.

Sequencing reactions were performed using both M13 Forward and M13 Reverse primers (included in the cloning kit) in a capillary DNA sequencer CEQ 8000 Genetic Analysis System (Beckman Coulter). The quality of the electropherograms was checked by eye in BioEdit 7.0.9.0. (Hall 1999).

A sequence-similarity search was performed in Blast to determine the similarities of the sequences obtained with other 5S rDNA sequences from DDBJ/EMBL/


A
B
alpha
beta
zeta
gamma
delta
epsilon
Ensis directus alpha
beta Ensis macha
zeta Ensis siliqua E. arcuatus E. ensis E. goreensis
gamma Ensis directus
delta Ensis directus E. arcuatus
epsilon Ensis macha
I
II
III
0.05

◀ **Fig. 1** Phylogenetic relationships of 5S rDNA variants in *Ensis* species reconstructed by means of a neighbor-joining phylogenetic tree. Six major clades, based mainly on differences in the NTS region, were identified and named using Greek letters. The tree is unrooted. **A** Decondensed tree with confidence values (P_B and P_C) shown at the internal nodes when $P_B \geq 50$. **B** Condensed tree displaying phylogenetic relationships between the six clades. Clades form three superclades designated I–III

GenBank. For sequence alignment we used ClustalX 2.08 (Larkin et al. 2007). Tandem repeats were characterized using Tandem Repeats Finder (Benson 1999).

All nucleotide sequence divergence and phylogenetic analyses were conducted in Mega 3.1 (Kumar et al. 2004). The extent of nucleotide sequence divergence was estimated by means of the Tamura–Nei distance (Tamura and Nei 1993), selecting the complete deletion option (i.e., not considering positions with gaps). Standard errors were calculated by the bootstrap option with 1000 replicates. Phylogenetic trees were constructed using the neighbor-joining method (Saitou and Nei 1987) based on the distance matrix. The reliability of the topologies was tested by the bootstrap test (Felsenstein 1985) and by the interior-branch test (Nei et al. 1985; Li 1989; Rzhetsky and Nei 1992) with 1000 replicates. For nucleotide diversity (π, Nei 1987) calculations we used DnaSP 4.0 (Rozas et al. 2003), excluding aligned gaps.

Results

We obtained 72 sequences experimentally. Sequence-similarity searches performed in Blast showed that all our sequences matched with other mollusc 5S rDNA available in the international nucleotide sequence databases, and identification of the 5S rRNA gene was performed by comparison with these sequences.

Many sequences corresponded to monomers formed by the last portion of the gene (88 bp), the NTS, and the first portion of the contiguous gene (32 bp). In two of the species analyzed (*E. directus* and *E. ensis*), we also obtained dimer sequences formed by two contiguous monomers. To maintain the similarity with other 5S rDNA sequences from the international nucleotide sequence databases, the last 32 bp of each sequence (monomers and dimers) was moved to the beginning. We added to our analysis 28 5S rDNA sequences belonging to the species *E. macha*, which were available in DDBJ/EMBL/GenBank, so we studied a dataset of 100 sequences (Table 1).

The 5S rRNA gene displayed 22 polymorphic sites in the *Ensis* species, and its size was 120 bp in all cases. The NTS showed a high degree of variation produced by several insertion–deletion polymorphisms (indels), nucleotide substitutions, and microsatellites (in 11 *E. directus* sequences and in one *E. macha* sequence). The size of the NTS region displayed high variation, between 286 and 619 bp.

Phylogenetic analysis of *Ensis* 5S rDNA revealed the existence of six well-supported clades (confidence values: for bootstrap test $P_B \geq 96\%$ and for interior-branch test $P_C \geq 87\%$) and three superclades ($P_B \geq 97\%$, $P_C \geq 98\%$) based on differences in the NTS region. Clades were named using Greek letters; alpha, beta, and zeta were given to superclade I, gamma and delta to superclade II, and epsilon to superclade III, based on mean distances between clades and phylogenetic

Table 1 Sequences of 5S ribosomal DNA studied in this work

Clade	Species	Accession number	*n*	π	SD
Alpha	*Ensis directus*	AM904878–AM904918*	41	0.0086	0.0012
Beta	*Ensis macha*	FM201452*; AM906171–AM906180; AM940998–AM941004	18	0.0152	0.0023
Gamma	*Ensis directus*	AM904919–AM904929*	11	0.0217	0.0034
Delta	*Ensis directus*	AM904930–AM904933*	4	0.0019	0.0007
	Ensis arcuatus	FM201453*; FM211693*	2	0	0
Epsilon	*Ensis macha*	AM906203–AM906208; AM941005–AM941009	11	0.0161	0.0017
Zeta	*Ensis siliqua*	FM201457–FM201462*; FM211689*	7	0.0059	0.0006
	Ensis arcuatus	FM201454–FM201456*	3	0.0238	0.0068
	Ensis ensis	FM211690–FM211691*	2	0.0460	0.0230
	Ensis goreensis	FM211692*	1	–	–

Accession numbers are from DDBJ/EMBL/GenBank. *n*, sample size; π, nucleotide diversity; SD, corresponding standard deviation

* Sequences obtained experimentally by the authors

clustering (Fig. 1, Tables 2 and 3). In the gene region, only one fixed mutation separated one clade from the others: a point mutation in position 12 (T → C) that characterized clade zeta.

Sequences of the species *E. directus* clustered in three clades (alpha, gamma, and delta) and in two superclades (I and II). The *E. macha* sequences clustered in beta (superclade I) and in epsilon (superclade III). Sequences from *E. arcuatus* clustered

Table 2 Mean sequence divergence[a] between and within clades of *Ensis* 5S rDNA

Clade	Alpha	Beta	Gamma	Delta	Epsilon	Zeta	Distance within clade	Standard error
Alpha	–	0.016	0.038	0.036	0.039	0.028	0.006	0.002
Beta	0.090	–	0.042	0.040	0.041	0.028	0.012	0.003
Gamma	0.330	0.370	–	0.012	0.048	0.048	0.012	0.003
Delta	0.296	0.337	0.055	–	0.048	0.047	0.004	0.002
Epsilon	0.348	0.365	0.459	0.455	–	0.043	0.012	0.003
Zeta	0.223	0.227	0.438	0.422	0.439	–	0.038	0.006

[a] Calculated using the Tamura–Nei distance. Below the diagonal, distance; above the diagonal, corresponding standard error

Table 3 Mean sequence divergence[a] between and within superclades of *Ensis* 5S rDNA

Superclade	I	II	III	Distance within superclade	Standard error
I	–	0.036	0.038	0.097	0.011
II	0.349	–	0.049	0.032	0.006
III	0.369	0.458	–	0.012	0.003

[a] Calculated using the Tamura–Nei distance. Below the diagonal, distance; above the diagonal, corresponding standard error

in zeta (superclade I) and in delta (superclade II). All *E. siliqua*, *E. ensis*, and *E. goreensis* sequences were in clade zeta (superclade I) (Fig. 1).

Tamura–Nei mean distances within clades were small (0.004–0.038) but larger between clades (0.055–0.459) (Table 2). In the same way, mean distances within superclades were small, but distance values between superclades were very large (0.349–0.458) (Table 3). Nucleotide diversity (π) was calculated by species within each of the six clades, resulting in small values of 0–0.0460 (Table 1), in comparison with π values obtained for other 5S rDNA bivalve sequences available in DDBJ/EMBL/GenBank (data not shown).

Regarding the organization of 5S rDNA dimers, we found that in the species *E. directus*, in some cases, sequences belonging to different clades were organized in tandem in the same clone. We sequenced 12 dimers alpha–alpha, but also one dimer alpha–gamma and two dimers alpha–delta. Contrarily, the dimer sequenced from *E. ensis* was composed of two zeta monomers.

Discussion

In the present study we found that several 5S rDNA variants, based on differences in the NTS region, occur in the genomes of *Ensis* razor shells. Three *Ensis* had more than one variant in their genomes, and the presence of more variants in the genomes of all the species cannot be discarded.

The 5S rDNA was used to design a genetic identification methodology (Fernández-Tajes and Méndez 2007) by means of PCR-RFLP, for some of the razor shells that we studied in the present work. These authors sequenced a small number of clones per species ($n = 2$), and sequences were not uploaded to DDBJ/EMBL/GenBank. Thus, we cannot compare their sequences with our own. The sequence sizes they provided, however, made us think they amplified only some of the 5S rDNA variants that we found, probably because their extension times in PCR (45 s) were shorter than ours, so shorter variants may have been favored in the amplification process.

The presence of several 5S rDNA variants has been reported in other bivalve molluscs, such as *Mytilus* mussels (Insua et al. 2001), the scallop *Aequipecten opercularis* (López-Piñón et al. 2008), and the cockle *Cerastoderma glaucum* (Freire et al. 2005), and in echinoderms (Caradonna et al. 2007), arthropods (Keller et al. 2006), and chordates (Martins and Galetti 2001; Daniels and Delany 2003). In other examples, 5S rDNA was found to be homogenized within the genome (Brown and Sugimoto 1974). Thus, the presence of multiple variants cannot be considered the rule, but given the examples listed above, neither is it the exception, especially because it was reported in four different animal phyla.

Nucleotide sequence divergence between the main *Ensis* 5S rDNA lineages (superclades I–III) is very large. This indicates that they split long ago, and probably were already present in the genome of the most common ancestor of these species (ancestral polymorphism). New sequence variants continued to emerge during evolution (in superclade I, alpha, beta, and zeta clades were differentiated, and the same happened in superclade II, with gamma and delta clades). The emergence of

new variants gives support to the idea that birth-and-death processes are responsible for the extant variation of this multigene family in *Ensis*.

Sequences of 5S rDNA belonging to different *Ensis* species clustered together in the same clades and superclades in the neighbor-joining tree. The lack of homogenization between 5S rDNA variants was evident from the analysis of distances between clades and superclades. Both features suggest again the involvement of birth-and-death processes in the long-term evolution of *Ensis* 5S rDNA. On the other hand, nucleotide diversity values calculated by species within each clade were rather small (e.g., $\pi = 0.0086$ in *E. directus* sequences from clade alpha). This suggests that homogenizing mechanisms are also taking part, reducing sequence divergence within each 5S rDNA variant in each species.

The presence of pseudogenes in a multigene family strongly suggests that the family evolves under a birth-and-death process (Rooney and Ward 2005). In *Ensis*, no evidence of pseudogenes was found. This is probably because both primers anneal in the 5S rRNA gene, and the presence of mutations in the gene would have considerably reduced the chance of amplification, cloning, and sequencing of pseudogenes.

The analysis of the 5S rRNA gene revealed only one diagnostic position (a cytosine in position 12) that was shared by all zeta sequences. Even though we found 22 polymorphic sites within the gene, all of them were point mutations (no indels in the gene) and could be selectively neutral. Purifying selection may have played a big role in maintaining the integrity of the 5S rRNA gene, as was found by Rooney and Ward (2005) and Fujiwara et al. (2009) in filamentous fungi and bitterlings, respectively.

Although the organization of 5S rDNA in *Ensis* chromosomes is still to be investigated, our data provide a very interesting feature in relation to *E. directus* sequences. Dimer sequences sampled from this species showed a striking organization: different 5S rDNA variants were organized in tandem in the same clone. The same has also been reported in two different fish groups, sturgeons (Robles et al. 2005) and bitterlings (Fujiwara et al. 2009), and it was suggested in gray mullets (Gornung et al. 2007). We cannot provide any certain explanation for the organization and the evolutionary processes that are maintaining this apparent organization of 5S rDNA in *E. directus*, until in situ hybridizations using specific probes are performed.

In short, (1) we have found several 5S rDNA variants in *Ensis* species, (2) 5S rDNA main lineages probably split a long time ago and may be a consequence of ancestral polymorphism, (3) new 5S rDNA sequence variants emerged during *Ensis* evolution, (4) phylogenetic analyses revealed a between-species clustering of *Ensis* 5S rDNA variants, (5) a lack of homogenization between variants was evident, (6) homogenizing mechanisms may be taking part within each variant in each species, and (7) scarce variation in the 5S rRNA gene was detected.

We conclude that birth-and-death processes are responsible for the extant variation of *Ensis* 5S rDNA. The low levels of variation found in the 5S rRNA gene are probably a consequence of selective pressures (i.e., purifying and positive selection). In filamentous fungi (Rooney and Ward 2005), sturgeons (Robles et al. 2005), and bitterlings (Fujiwara et al. 2009), the evolution of 5S rDNA was reported

to be a consequence of several evolutionary processes. In the same way, the long-term evolution of *Ensis* 5S rDNA seems to be driven by a combination of birth-and-death processes and selection.

Acknowledgements We thank K. Thomas Jensen, Anne S. Lousdal, Ana de la Torriente, Rudo von Cosel, Virginie Héros, and Barbara Buge for providing us with some of the specimens studied. We are in debt to Ángeles Cid for her invaluable help. The support of the Consellería de Educación e Ordenación Universitaria (Xunta de Galicia, Spain) is greatly appreciated.

References

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Brown DD, Sugimoto K (1974) The structure and evolution of ribosomal and 5S DNAs in *Xenopus laevis* and *Xenopus mulleri*. Cold Spring Harb Symp Quant Biol 38:501–505

Caradonna F, Bellavia D, Clemente AM, Sisino G, Barbieri R (2007) Chromosomal localization and molecular characterization of three different 5S ribosomal DNA clusters in the sea urchin *Paracentrotus lividus*. Genome 50:867–870

Daniels LM, Delany ME (2003) Molecular and cytogenetic organization of the 5S ribosomal DNA array in chicken (*Gallus gallus*). Chromosome Res 11:305–317

Danna KJ, Workman R, Coryell V, Keim P (1996) 5S rRNA genes in tribe Phaseoleae: array size, number, and dynamics. Genome 39:445–455

Eickbush TH, Eickbush DG (2007) Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics 175:477–485

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Fernández-Tajes J, Méndez J (2007) Identification of the razor clam species *Ensis arcuatus*, *E. siliqua*, *E. directus*, *E. macha*, and *Solen marginatus* using PCR-RFLP analysis of the 5S rDNA region. J Agric Food Chem 55:7278–7282

Freire R, Insua A, Méndez J (2005) *Cerastoderma glaucum* 5S ribosomal DNA: characterization of the repeat unit, divergence with respect to *Cerastoderma edule*, and PCR-RFLPs for the identification of both cockles. Genome 48:427–442

Fujiwara M, Inafuku J, Takeda A, Watanabe A, Fujiwara A, Kohno S, Kubota S (2009) Molecular organization of 5S rDNA in bitterlings (Cyprinidae). Genetica 135:355–365

Giegerich R, Meyer F, Schleiermacher C (1996) GeneFisher: software support for the detection of postulated genes. Proc Int Conf Intell Syst Mol Biol 4:68–77

Gornung E, Colangelo P, Annesi F (2007) 5S ribosomal RNA genes in six species of Mediterranean grey mullets: genomic organization and phylogenetic inference. Genome 50:787–795

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Insua A, Freire R, Ríos J, Méndez J (2001) The 5S rDNA of mussels *Mytilus galloprovincialis* and *M. edulis*: sequence variation and chromosomal location. Chromosome Res 9:495–505

Keller I, Chintauan-Marquier IC, Veltsos P, Nichols RA (2006) Ribosomal DNA in the grasshopper *Podisma pedestris*: escape from concerted evolution. Genetics 174:863–874

Kumar S, Tamura K, Nei M (2004) Mega3: integrated software for molecular evolutionary genetic analysis and sequence alignment. Brief Bioinform 5:150–163

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Leo NP, Barker SC (2002) Intragenomic variation in ITS2 rDNA in the louse of humans, *Pediculus humanus*: ITS2 is not a suitable marker for population studies in this species. Insect Mol Biol 11:651–657

Li WH (1989) A statistical test of phylogenies estimated from sequence data. Mol Biol Evol 6:424–435

Li Z, Huang S, Jin W, Ning S, Song Y, Li L (2002) Determination of copy number for 5S rDNA and centromeric sequence RCS2 in rice by Fiber-FISH. Chin Sci Bull 47:214–217

Little RD, Braaten BC (1989) Genomic organization of human 5S rDNA and sequence of one tandem repeat. Genomics 4:376–383

López-Piñón MJ, Freire R, Insua A, Méndez J (2008) Sequence characterization and phylogenetic analysis of the 5S ribosomal DNA in some scallops (Bivalvia: Pectinidae). Hereditas 145:9–19

Martins C, Galetti PM (2001) Two 5S rDNA arrays in Neotropical fish species: is it a general rule for fishes? Genetica 111:439–446

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Hughes AL (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: Tsuji K, Aizawa M, Sasazuki T (eds) 11th Histocompatibility workshop and conference. Oxford University Press, Oxford (UK), pp 27–38

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152

Nei M, Stephens JC, Saitou N (1985) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol Biol Evol 2:66–85

Robles F, de la Herrán R, Ludwig A, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA (2005) Genomic organization and evolution of the 5S ribosomal DNA in the ancient fish sturgeon. Genome 48:18–28

Rooney AP (2004) Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans. Mol Biol Evol 21:1704–1711

Rooney AP, Ward TJ (2005) Evolution of large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proc Natl Acad Sci USA 102:5084–5098

Rozas J, Sánchez-Del Barrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sword GA, Senior LB, Gaskin JF, Joern A (2007) Double trouble for grasshopper molecular systematics: intra-individual heterogeneity of both mitochondrial 12S-valine-16S and nuclear internal transcribed spacer ribosomal DNA sequences in *Hesperotettix viridis* (Orthoptera: Acrididae). Syst Entomol 32:420–428

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S et al (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature 415:871–880

10.2 Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers

Joaquín Vierna, Andrés Martínez-Lage, Ana M. González-Tizón (2010) Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers. Genome 53:23-34.

Bibliometrics 2012 JCR Science Edition

Impact factor: 1.668

Biotechnology & Applied Microbiology: Q3

Genetics & Heredity: Q4

Analysis of ITS1 and ITS2 sequences in *Ensis* razor shells: suitability as molecular markers at the population and species levels, and evolution of these ribosomal DNA spacers

Joaquín Vierna, Andrés Martínez-Lage, and Ana M. González-Tizón

Abstract: Internal transcribed spacer 1 and 2 (ITS1 and ITS2) sequences were analysed in *Ensis* razor shells (Mollusca: Bivalvia: Pharidae). We aimed to (1) test ITS1 and ITS2 as molecular markers at the population level in the successful alien *E. directus* (Conrad, 1843); (2) test these spacers at the species level in *E. directus* and three other *Ensis* species, *E. siliqua* (L., 1758), *E. macha* (Molina, 1782), and *E. magnus* (Schumacher, 1817); and (3) analyse the evolutionary processes that may be shaping *Ensis* ITS1 and ITS2 extant variation. In *E. directus*, despite the intragenomic divergence detected, ITS1 and ITS2 were informative in differentiating the geographic areas considered (Denmark and Canada) by means of both the insertion-deletion polymorphism and the nucleotide polymorphism. In this species, the 5.8S ribosomal gene (5.8S) showed scarce polymorphism. At the species level, maximum parsimony and maximum likelihood analyses revealed that ITS1 and ITS2 may be suitable to reconstruct *Ensis* phylogenetic relationships. Finally, the evolutionary models that best fit the long-term evolution of *Ensis* ITS1–5.8S–ITS2 are discussed. A mixed process of concerted evolution, birth-and-death evolution, and selection is chosen as an option that may reconcile the long-term evolution of *Ensis* ITS1–5.8S–ITS2 and 5S ribosomal DNA.

Key words: internal transcribed spacers, rDNA, concerted evolution, birth-and-death evolution, mixed process, indel polymorphism.

Résumé : Les séquences des espaceurs internes transcrits 1 et 2 (ITS1 et ITS2) ont été analysées chez les couteaux du genre *Ensis* (Mollusca ; Bivalvia : Pharidae). Les auteurs voulaient (1) évaluer ITS1 et ITS2 en tant que marqueurs moléculaires au niveau des populations chez l'espèce introduite *E. directus* (Conrad, 1843) ; (2) évaluer ces marqueurs au niveau de l'espèce en incluant trois autres espèces d'*Ensis*, *E. siliqua* (L., 1758), *E. macha* (Molina, 1782) et *E. magnus* (Schumacher, 1817) ; et (3) analyser les processus évolutifs qui façonneraient la variation observée chez les ITS1 et ITS2 au sein du genre *Ensis*. Chez l'*E. directus*, en dépit de la divergence intragénomique observée, ITS1 et ITS2 étaient utiles pour distinguer les zones géographiques à l'étude (le Danemark et le Canada) sur la base du polymorphisme tant de type insertion-délétion que nucléotidique. Chez cette espèce, le gène d'ARN ribosomique 5,8S (5,8S) affichait peu de polymorphisme. Au niveau de l'espèce, des analyses de parcimonie maximale et de vraisemblance maximale ont révélé que ITS1 et ITS2 seraient possiblement utiles pour établir les relations phylogénétiques chez le genre *Ensis*. Finalement, les auteurs discutent des modèles évolutifs qui concordent le mieux avec l'évolution à long terme des séquences ITS1–5,8S–ITS2 chez le genre *Ensis*. Un processus mixte combinant l'évolution de type naissance-mort (« birth-and-death ») et la sélection est retenu comme une option pouvant concilier l'évolution à long terme des séquences ITS1–5,8S–ITS2 et 5S chez le genre *Ensis*.

Mots-clés : espaceurs internes transcrits, ADNr, évolution concertée, évolution de type naissance-mort, processus mixte, polymorphisme de type indel.

[Traduit par la Rédaction]

Introduction

The evolution of nuclear ribosomal DNA (nrDNA) is traditionally explained by concerted evolution. Homogenizing mechanisms such as unequal crossovers and gene conversions eventually maintain the sequence similarity among all repeats of the array, and even among repeats belonging to different arrays (for a review see Eickbush and Eickbush 2007).

Nevertheless, it has recently been demostrated that birth-

Received 10 August 2009. Accepted 22 October 2009. Published on the NRC Research Press Web site at genome.nrc.ca on 15 December 2009.

Corresponding Editor: A. Civetta.

J. Vierna, A. Martínez-Lage, and A.M. González-Tizón.[1] Department of Molecular and Cell Biology, Evolutionary Biology Group (GIBE), Universidade da Coruña, A Zapateira s/n, E-15071 La Coruña, Spain.

[1]Corresponding author (e-mail: hakuna@udc.es).

and-death processes may be responsible, at least in part, for the evolution of nrDNA in some organisms, including *Ensis* razor shells (Mollusca: Bivalvia: Pharidae) (Rooney 2004; Rooney and Ward 2005; Fujiwara et al. 2009; Vierna et al. 2009). Under a birth-and-death model, gene duplications generate new genes that can persist in the genome for long periods, degenerate into pseudogenes, or become deleted. In a phylogenetic analysis of genes from several closely related taxa, sequences will show a between-species clustering pattern, in contrast to the within-species clustering pattern that is expected under concerted evolution. However, there are two exceptions to this rule: (1) recent gene duplicates and (2) multigene families that experience rapid gene turnover due to birth-and-death evolution (e.g., Zhang et al. 2000). In the latter case, rapid gene turnover can lead to the creation of species-specific gene clusters as a result of frequent gene duplication and loss (Rooney 2004).

Natural selection is expected to play a role in the long-term evolution of nuclear ribosomal genes, under either a concerted or birth-and-death evolution scenario, reducing sequence variation in the coding regions.

Under a strict concerted evolution scenario, intragenomic divergence remains very low, and all copies of a nrDNA gene or spacer from a single genome can be considered a single locus in phylogenetic or molecular population studies. Under birth-and-death evolution, substantially different copies may exist in the genome (except in the two cases stated above), therefore increasing the levels of intragenomic divergence. Since intragenomic variants produced by birth-and-death processes are paralogous, researchers must be very careful when using nrDNA genes or spacers as molecular markers at any taxonomic level. Comparisons between paralogous sequences may completely skew the results of the survey.

Even though concerted evolution functions extremely well on all nrDNA genes and spacers in the vast majority of organisms (Eickbush and Eickbush 2007), some authors have recommended analysing the intragenomic divergence levels before using the internal transcribed spacers of the major ribosomal genes (ITS1 and ITS2) as molecular markers (Leo and Barker 2002; Wörheide et al. 2004; Alvarez et al. 2007; Sword et al. 2007). Interestingly, according to Harris and Crandall (2000), in all cases reported so far where intragenomic divergence was looked for within ITS1 and ITS2, it was found to some degree.

In the present work we analysed sequence variation within ITS1 and ITS2 (hereafter, ITS) in *Ensis* razor shells. We studied their suitability as molecular markers, and the evolutionary processes that may be responsible for the long-term evolution of *Ensis* ITS1–5.8S–ITS2. We considered four *Ensis* species: *E. directus* (Conrad, 1843) (syn. *E. americanus* (Gould, 1870)), which is native to the Atlantic coast of North America but was introduced in European waters at the end of the 1970s (Cosel et al. 1982); *E. siliqua* (L., 1758) and *E. magnus* (Schumacher, 1817) (syn. *E. arcuatus* var. *ensoides* Van Urk, 1964 and *E. arcuatus* var. *norvegica* Van Urk, 1964), which are native to Atlantic Europe; and *E. macha* (Molina, 1782), which occurs along the southern coasts of Chile and Argentina. Taking into account the levels of intragenomic divergence detected, ITS were studied as molecular markers at the population level in *E. directus* and at the species level in all four species.

Material and methods

Ensis directus individuals were sampled at three Danish localities (Sillerslev, Sundsøre, and Juvre Deep) and one Canadian locality (Long Pond, Conception Bay, Newfoundland). All individuals were identified as *E. directus* according to morphological characters including animal size, shape and length of muscle scars, and shell colour and curvature. In the present article, all taxon names follow Cosel (2009) (when applicable). Animals were stored in 100% ethanol until DNA extraction.

Extraction of genomic DNA was performed from foot tissue using the NucleoSpin Tissue kit (Macherey-Nagel GmbH and Co. KG). For PCR we used a pair of primers that anneal at the 3′ end of the 18S ribosomal gene and the 5′ end of the 28S ribosomal gene (Heath et al. 1995) that were previously used on several bivalve species (Fernández et al. 2001; Insua et al. 2003; Toro et al. 2004; Mahidol et al. 2007). ITS1, the 5.8S ribosomal gene (hereafter, 5.8S), and ITS2 were amplified. Each reaction was performed in a total volume of 25 μL containing ~25 ng of genomic DNA, 0.625 U of *Taq* DNA polymerase (Roche Diagnostics), 5 nmol of each dNTP (Roche Diagnostics), 20 pmol of each primer, and the buffer recommended by the polymerase supplier. PCR conditions were as follows: an initial denaturation step at 94 °C for 3 min followed by 35 cycles of denaturation at 94 °C for 20 s, annealing at 54 °C for 20 s, and extension at 72 °C for 45 s, and a final extension at 72 °C for 5 min. PCR products were run on agarose gels, stained with ethidium bromide, and imaged under UV light. A single band of about 1000 bp was obtained from all individuals analysed, and it was cloned using the TOPO TA Cloning Kit (Invitrogen). Several transformant colonies (n = 2–10) from each individual were selected at random and grown in LB medium. A QIAprep Spin Miniprep Kit (QIAGEN) was used to purify the plasmids, which were sequenced using both M13 Forward and M13 Reverse primers (included in the cloning kit) in a capillary DNA sequencer (CEQ 8000 Genetic Analysis System, Beckman Coulter). Electropherograms were examined and assembled in BioEdit 7.0.9.0 (Hall 1999). A sequence similarity search was performed in BLAST (http://blast.ncbi.nlm.nih.gov) to determine the similarities of the sequences obtained to other ITS1–5.8S–ITS2 sequences from DDBJ/EMBL/GenBank. BLAST confirmed that all sequences obtained were similar to other mollusc ITS1–5.8S–ITS2 sequences, and permitted identification of the flanking regions of each ITS. Sequences were deposited in the DDBJ/EMBL/GenBank databases under the following accession numbers: FN391027–FN391081 for ITS1 and FN391082–FN391136 for ITS2.

To examine the potential of both ITS as phylogenetic markers within genus *Ensis*, we downloaded sequences from *E. siliqua* (accession Nos. AJ966667–AJ966681), *E. macha* (AM933624–AM933631), and *E. magnus* (AJ966682–AJ966697) available in DDBJ/EMBL/GenBank. No 5.8S sequences from these species were available.

Sequence alignments were performed in ClustalX 2.08 (Larkin et al. 2007) (penalties for gap opening = 7 and gap

Table 1. Number of sequences sampled (*n*), number of sequencetypes obtained (*s*), and haplotype diversities (Hd) of *Ensis directus* internal transcribed spacers (ITS1 and ITS2).

			ITS1		ITS2	
		n	*s*	Hd (mean ± SE)	*s*	Hd (mean ± SE)
sp	*Ensis directus*	55	25	0.921±0.024	19	0.931±0.013
area	Denmark	33	19	0.939±0.026	14	0.907±0.028
area/loc	Canada / Long Pond	22	6	0.632±0.104	7	0.771±0.045
loc	Sillerslev	6	5	0.933±0.122	3	0.733±0.155
loc	Sundsøre	19	9	0.848±0.068	8	0.836±0.065
loc	Juvre Deep	8	6	0.893±0.111	6	0.893±0.111
ind	Long Pond 12	7	1	0	1	0
ind	Long Pond 13	6	2	0.333±0.215	2	0.333±0.215
ind	Long Pond 15	3	2	0.667±0.314	1	0
ind	Long Pond 16	3	2	0.667±0.314	2	0.667±0.314
ind	Long Pond 17	3	2	0.667±0.314	2	0.667±0.314
ind	Sillerslev 3	3	3	1.000±0.272	2	0.667±0.314
ind	Sillerslev 8	3	2	0.667±0.314	1	0
ind	Sundsøre 53	3	2	0.667±0.314	1	0
ind	Sundsøre 58	10	4	0.533±0.180	4	0.533±0.180
ind	Sundsøre 66	3	1	0	2	0.667±0.314
ind	Sundsøre 67	3	2	0.667±0.314	1	0
ind	Juvre Deep 103	3	1	0	1	0
ind	Juvre Deep 117	3	3	1.000±0.272	3	0.074±0.272
ind	Juvre Deep 119	2	2	1.000±0.500	2	1.000±0.500

Note: Sequences were grouped in several categories according to the species (sp), area, locality (loc), and individual (ind) they belong to.

extension = 3) and manually adjusted for local optimization in MEGA 4.0.2 (Tamura et al. 2007). To use as many nucleotide positions as possible in the subsequent analyses, we performed three different alignments for each ITS: one with all *E. directus* sequences, one with *E. directus* and *E. macha* sequences, and one with all sequences from the four *Ensis* species. In this case, as some regions did not align properly, the alignment was corrected in the Gblocks Server (http://molevol.cmima.csic.es/castresana/Gblocks_server.html) (Castresana 2000) to increase the probability of considering only homologous positions in the analyses. All alignments are available as supplementary material (Figs. S1–S6).[2]

Sequence lengths and mean nucleotide compositions were obtained from MEGA 4.0.2 (Tamura et al. 2007). Using DnaSP 4.9 (Rozas et al. 2003), we studied the insertion-deletion (indel) polymorphism of each marker. To analyse the potential of indel polymorphism to differentiate sequences from different sampling localities and geographic areas, a nexus file with the indel polymorphism of both ITS was generated in DnaSP 4.9 (Rozas et al. 2003) and run in Paup*4.0b10 (Swofford 2002) under maximum parsimony (MP) (settings were those used in the phylogenetic analysis of *E. siliqua*, *E. macha*, *E. directus*, and *E. magnus* ITS; see below). The number of "sequencetypes" and the haplotype diversities were also calculated in DnaSP 4.9 (Rozas et al. 2003), excluding nucleotide positions with gaps. In this study we use the word "sequencetype" to denote a single and unique type of ITS, following Wörheide et al. (2002). Mean sequence divergences within and between groups and the maximum divergence between sequences were obtained by means of the *p*-distance using MEGA 4.0.2 (Tamura et al. 2007). To study a possible clustering of *E. directus* ITS sequences according to sampling localities or geographic areas, maximum likelihood (ML) phylogenetic trees were reconstructed using the PhyML 3.0 Web server (http://www.atgc-montpellier.fr/phyml/) (Guindon and Gascuel 2003) for each ITS and for a concatenated data set of both ITS. The best-fit model of nucleotide substitution for each data set was selected by statistical comparison of 88 different models using jModelTest 0.1.1 (Posada 2008) and applying the Akaike information criterion. Starting trees were obtained using the BIONJ algorithm (Gascuel 1997) and gaps were treated as unknown characters. Tree topologies were estimated by a nearest neighbor interchange (Jarvis et al. 1983 and references therein). Node support was estimated by the bootstrap test (Felsenstein 1985) (100 replicates).

Ensis phylogenetic trees were reconstructed using Gblocks alignments (Figs. S3–S4) and two different phylogenetic methods, ML and MP. To save computational time, duplicate sequences were deleted. ML analyses were performed as for *E. directus* sequences (see above). For MP analyses, a heuristic search was conducted in Paup*4.0b10 (Swofford 2002) using the tree-bisection-reconnection (TBR) branch-swapping algorithm, and assuming equal weights and unordered character states for all characters. Gaps were treated as a fifth character state or as missing in-

[2] Supplementary data for this article are available on the journal Web site (http://genome.nrc.ca) or may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Building M-55, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada. DUD 5311. For more information on obtaining material refer to http://cisti-icist.nrc-cnrc.gc.ca/eng/ibp/cisti/collection/unpublished-data.html.

Fig. 1. Maximum parsimony majority-rule consensus cladogram constructed on the basis of the insertion-deletion polymorphism of the internal transcribed spacers (ITS1 and ITS2) of *Ensis directus* (concatenated data set). Bootstrap values ≥ 50% are indicated at the nodes. Sampling localities, individuals, and clones are specified.

formation. Starting trees were obtained via stepwise addition with random addition of sequences (10 replicates). The number of trees held at each step during stepwise addition was one. In both ML and MP analyses, node support was estimated by the bootstrap test (Felsenstein 1985) with 1000 replicates.

All MP and ML phylogenetic trees were edited in FigTree 1.2.2 (Andrew Rambaut; http://tree.bio.ed.ac.uk/software/figtree/).

Results

Ensis directus ITS

A total of 55 ITS1–5.8S–ITS2 sequences were obtained experimentally from *E. directus* (the number of clones per individual and the number of individuals per geographic area or locality are summarized in Table 1). Average GC contents were 58.9% for ITS1, 63% for ITS2, and 57.4% for the 5.8S. The 5.8S displayed no indels and only 4 point mutations (positions 17A>G, 80A>G, 101T>C, and 135A>G).

The length of ITS1 ranged between 484 and 510 bp, whereas the length of ITS2 was less variable (295–299 bp). The 5.8S was 157 bp in all clones. ITS1 length from Danish sequences ranged between 497 and 510 bp, and for ITS2 the length was 295–299 bp. In Canadian sequences, ITS length variation was low: 484–486 bp for ITS1 and 296 bp (no length variation) for ITS2. Intragenomic length variation in ITS1 was present in 4 individuals from Denmark (Sillerslev 3, Sundsøre 53, Sundsøre 58, and Sundsøre 67) and 2 Canadian individuals (Long Pond 13 and Long Pond 17). Three Danish animals displayed intragenomic length variation in ITS2 (Sundsøre 58, Juvre Deep 117, and Juvre Deep 119).

The *E. directus* ITS1 alignment (Fig. S1) contained 16 indel events, with an average indel event length of 2.625 bp. In contrast with ITS1, the ITS2 alignment (Fig. S2) contained only 3 indel events, with an average indel event length of 1.667 bp. The analysis of indel polymorphism in each geographic area showed that for ITS1, in Danish sequences there were 13 indel events, with an average indel event length of 1.692 bp, whereas in Canadian sequences there were only 3 indel events (average indel event length = 2 bp). In Danish ITS2 sequences, there were 2 indel events, with an average indel event length of 2 bp, whereas Canadian ITS2 sequences displayed no indel polymorphism. A thorough examination of *E. directus* ITS alignments showed that some indels were characteristic of one of the two geographic areas under study. For example, in ITS1, a TTG insertion in position 178 and a CGAGACGGCGTTAAC deletion in position 236 characterized individuals from Canada. In ITS2, a single nucleotide deletion in position 275 also characterized all Canadian individuals. The MP cladogram constructed on the basis of ITS indel polymorphism (Fig. 1) showed that sequences belonging to each geographic area grouped together with a bootstrap support of 88%, but no resolution according to sampling localities was obtained.

The 55 ITS1 and ITS2 sequences sampled from *E. directus* produced 25 ITS1 sequencetypes and 19 ITS2 sequencetypes. In both markers, haplotype diversity was higher in Denmark than in Canada (Table 1). Of the 14 individuals analysed, only 3 had a single ITS1 sequencetype and only 6 had a single ITS2 sequencetype. There were no shared sequencetypes between Denmark and Canada.

Some authors prefer to use gap information when calculating sequence divergences, whereas others do not. Therefore, we show all mean divergences in Tables 2–3, though *p*-distance values did not substantially differ when positions with gaps were considered or excluded. Within groups, *p*-distance values did not decrease from species to individuals in either ITS. In fact, in some cases the mean divergence within individuals was of the same order of magnitude as the mean divergence within areas. For example, for ITS1, individual Sundsøre 67 showed a mean intragenomic divergence of 0.018 ± 0.005, while the value for all Danish sequences was 0.013 ± 0.003 (Table 2). ITS1 mean intragenomic divergence was particularly high in individuals Sundsøre 67 and Juvre Deep 119.

Alignments revealed some divergent copies of ITS1 and ITS2 in the genome of *E. directus*. One ITS1 clone from individual Sundsøre 67 displayed 14 point mutations and 1 indel compared with the other clones of this individual. Similarly, one clone from Juvre Deep 119 displayed 5 point mutations and 3 indels. That clone from Sundsøre 67, however, did not show any variation in its ITS2 region, although variation would be expected because of the linkage of ITS1 and ITS2 sequences in the same clone. One ITS2 clone from individual Juvre Deep 119 displayed 7 point mutations and 1 indel, making its mean divergence quite high (0.024 ± 0.009). All ITS2 clones from individuals Juvre Deep 103 and Sillerslev 8 and one ITS2 clone from Juvre Deep 119 shared some point mutations between positions 269–276. This signal was strong enough to group all these sequences in one clade in the ITS phylogenetic tree (Fig. 2). In pairwise comparisons (matrix not shown), the maximum *p*-distance obtained for ITS1 was 0.052 ± 0.010, between one clone from individual Sundsøre 67 and two clones belonging to individuals Sillerslev 8 and Long Pond 13. For ITS2, the maximum *p*-distance obtained was 0.034 ± 0.011, between one clone from Sundsøre 58 and one clone from Sillerslev 8.

ITS sequences did not cluster according to sampling localities in the area of Denmark in the ML trees. Instead, sequences from different localities were intermixed together in the same clades. However, a clustering by geographic areas was found in all trees. Higher bootstrap support (100%) for each geographic clade was obtained with the concatenated set of data (both ITS) (Fig. 2). ITS1 sequences alone were also informative in differentiating each geographic clade (bootstrap = 100%, tree not shown). ITS2 sequences were able to differentiate each geographic clade, but in this case bootstrap support was low (bootstrap = 43%, tree not shown).

The mean sequence divergence between *E. directus* local-

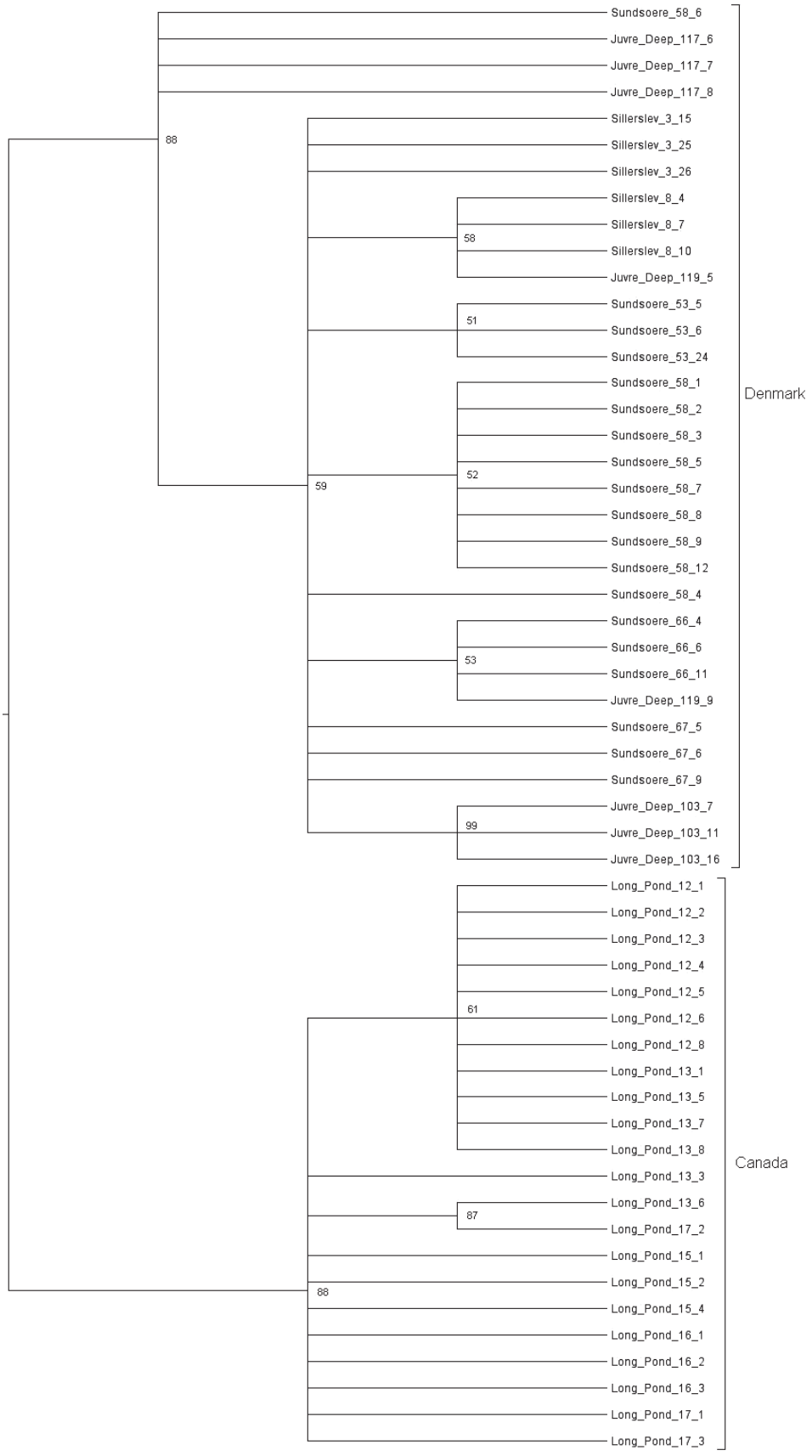

Sundsoere_58_6
Juvre_Deep_117_6
Juvre_Deep_117_7
Juvre_Deep_117_8
Sillerslev_3_15
Sillerslev_3_25
Sillerslev_3_26
Sillerslev_8_4
Sillerslev_8_7
Sillerslev_8_10
Juvre_Deep_119_5
Sundsoere_53_5
Sundsoere_53_6
Sundsoere_53_24
Sundsoere_58_1
Sundsoere_58_2
Sundsoere_58_3
Sundsoere_58_5
Sundsoere_58_7
Sundsoere_58_8
Sundsoere_58_9
Sundsoere_58_12
Sundsoere_58_4
Sundsoere_66_4
Sundsoere_66_6
Sundsoere_66_11
Juvre_Deep_119_9
Sundsoere_67_5
Sundsoere_67_6
Sundsoere_67_9
Juvre_Deep_103_7
Juvre_Deep_103_11
Juvre_Deep_103_16
Denmark
Long_Pond_12_1
Long_Pond_12_2
Long_Pond_12_3
Long_Pond_12_4
Long_Pond_12_5
Long_Pond_12_6
Long_Pond_12_8
Long_Pond_13_1
Long_Pond_13_5
Long_Pond_13_7
Long_Pond_13_8
Long_Pond_13_3
Long_Pond_13_6
Long_Pond_17_2
Long_Pond_15_1
Long_Pond_15_2
Long_Pond_15_4
Long_Pond_16_1
Long_Pond_16_2
Long_Pond_16_3
Long_Pond_17_1
Long_Pond_17_3
Canada
88
58
51
59
52
53
99
61
87
88

Table 2. Mean sequence divergence (*p*-distance, mean ± SE) within groups (species, areas, localities, and individuals) for both internal transcribed spacers (ITS1 and ITS2) of *Ensis directus*.

			ITS1		ITS2	
		n	Gaps considered	Gaps excluded	Gaps considered	Gaps excluded
sp	*E. directus*	55	0.020±0.004	0.018±0.004	0.013±0.004	0.013±0.004
area	Denmark	33	0.014±0.003	0.013±0.003	0.014±0.004	0.013±0.004
area/loc	Canada/Long Pond	22	0.002±0.001	0.002±0.001	0.005±0.002	0.005±0.002
loc	Sillerslev	6	0.016±0.004	0.013±0.004	0.013±0.005	0.013±0.005
loc	Sundsøre	19	0.012±0.003	0.012±0.003	0.006±0.003	0.006±0.003
loc	Juvre Deep	8	0.008±0.003	0.009±0.002	0.018±0.006	0.017±0.005
ind	Long Pond 12	7	0	0	0	0
ind	Long Pond 13	6	0.002±0.001	0.002±0.001	0.001±0.001	0.001±0.001
ind	Long Pond 15	3	0.001±0.001	0.001±0.001	0	0
ind	Long Pond 16	3	0.004±0.002	0.004±0.002	0.002±0.002	0.002±0.002
ind	Long Pond 17	3	0.001±0.001	0.001±0.001	0.002±0.002	0.002±0.002
ind	Sillerslev 3	3	0.004±0.002	0.003±0.002	0.002±0.002	0.002±0.002
ind	Sillerslev 8	3	0.003±0.002	0.003±0.002	0	0
ind	Sundsøre 53	3	0.001±0.001	0.001±0.001	0	0
ind	Sundsøre 58	10	0.006±0.002	0.006±0.002	0.003±0.001	0.003±0.001
ind	Sundsøre 66	3	0	0	0.002±0.002	0.002±0.002
ind	Sundsøre 67	3	0.017±0.005	0.018±0.005	0	0
ind	Juvre Deep 103	3	0	0	0	0
ind	Juvre Deep 117	3	0.003±0.002	0.003±0.002	0.007±0	0.007±0
ind	Juvre Deep 119	2	0.012±0.005	0.012±0.005	0.023±0.008	0.024±0.009

Note: Mean sequence divergence was calculated with gaps considered and with gaps excluded; the latter values are cited in the text.

ities was compared with the mean sequence divergence between different *Ensis* species (Table 3). In all cases except one, *p*-distances between localities were smaller than *p*-distances between species, even when the species alignment was corrected in the Gblocks Server (Castresana 2000) (Figs. S3–S4) (after alignment correction, mean divergence between species is somewhat reduced). The exception is the mean sequence divergence between *E. magnus* and *E. siliqua* (0.019 ± 0.006), which was somewhat smaller than the mean sequence divergence between Long Pond and the Danish localities in the ITS1 analysis. According to *p*-distance, *E. macha* was the closest species to *E. directus* (for ITS1, 0.093 ± 0.013; for ITS2, 0.109 ± 0.018, using the Gblocks alignment). The *p*-distances between the Canadian locality and the Danish localities ranged between 0.025 ± 0.006 and 0.028 ± 0.006 for ITS1 and between 0.013 ± 0.005 and 0.021 ± 0.006 for ITS2.

Ensis ITS phylogenetic analyses

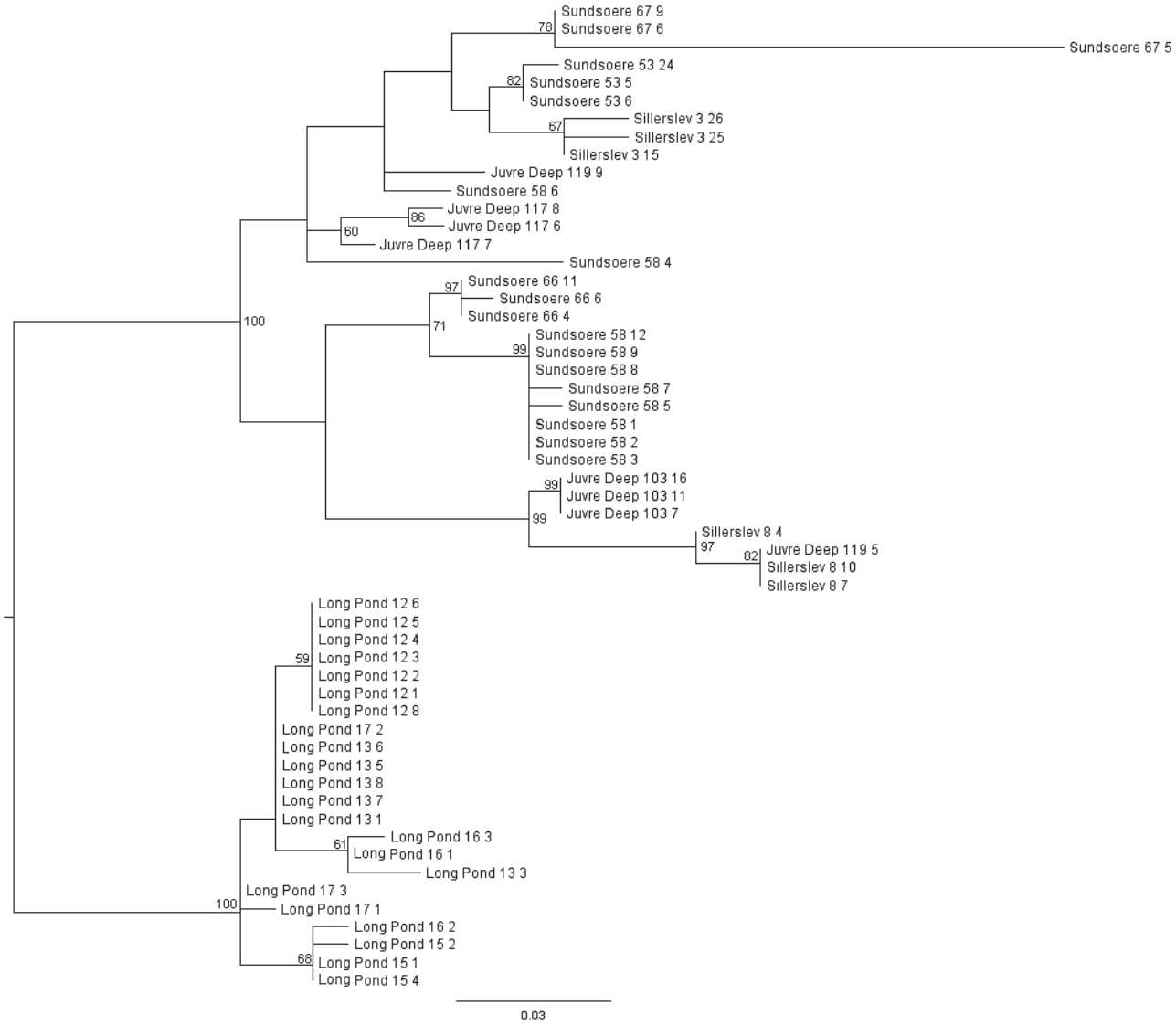
In the ITS1 ML analysis, all sequences belonging to the same species grouped together, with the exception of *E. siliqua*. In this species, ITS1 sequences did not form a single clade in the phylogenetic tree (Fig. 3*a*). However, in the ITS2 ML analysis, all sequences belonging to a single species were recovered as monophyletic, and species bootstrap support was higher than for ITS1 (Fig. 3*b*). Under MP, sequences belonging to a single species also grouped together, and bootstrap support for species nodes was higher with ITS2 sequences (≥97%). In MP analyses, gaps were treated as a fifth character state or as missing information. When gaps were treated as a fifth character state in the ITS1 MP analysis, the species node support increased in *E. siliqua*, decreased in *E. magnus*, and did not change in *E. directus* and *E. macha*. In the ITS2 MP analysis, species node support was higher when gaps were considered. The resulting topologies under MP criteria were identical regardless of the way in which gaps were treated.

Discussion

Comparison of *Ensis directus* sequences with other mollusc ITS sequences

Marine molluscs are a diverse group according to the length and GC content of ITS1 (Chow et al. 2009). These authors studied the ITS1 of several marine animals, and extensive data regarding length and GC content for marine mollusc species are available. Average ITS1 length was 492.4 bp, and GC content was 55.9%. The values we obtained for *E. directus* ITS1 sequences are very similar to those values, so they are not an exception. Data on ITS2 length and GC content in Mollusca are not abundant, but some examples for bivalve species are available. ITS2 length was only 60 bp in the crocus giant clam, *Tridacna crocea* (Yu et al. 2000). It ranged between 317 and 446 bp in the Unionoidea species *Unio pictorum*, *U. tumidus*, *U. crassus*, *Anodonta anatina*, *A. cygnea*, *Pseudanodonta complanata*, and *Margaritifera margaritifera* (Källersjö et al. 2005). In the four scallop species studied by Insua et al. (2003), *Aequipecten opercularis*, *Mimachlamys varia*, *Hinnites distortus*, and *Pecten maximus*, ITS2 length ranged between 270 and 294 bp and GC content was 47%–50%. In the Veneridae species *Meretrix meretrix*, *Cyclina sinensis*, *Mercenaria mercenaria*, and *Protothaca jedoensis*, ITS2 length was 281–412 bp and GC content was 65.21%–67.87% (Cheng et al. 2006). Thus, *E. directus* ITS2 lengths were similar to those found in the scallop and venerid species. ITS2 GC content in *E. directus* is similar to that in venerids, as would be expected considering that scallops are

Fig. 2. Maximum likelihood majority-rule consensus tree of *Ensis directus* internal transcribed spacers (ITS1 and ITS2) (concatenated data set) reconstructed following the GTR+I+G model. Bootstrap values ≥ 50% are indicated at the nodes. Sampling localities, individuals, and clones are specified.



Pteriomorphia bivalves, and venerids and *Ensis* species are Heteroconchia.

ITS as molecular markers in *Ensis directus*

Analysis of ITS variation in *E. directus* was performed by studying the length, indel polymorphism, haplotype diversity, and mean sequence divergence (*p*-distance) of each ITS at three different levels: individuals, localities, and geographic areas. Danish sequences were more variable than Canadian sequences in length, and they displayed a greater indel polymorphism, a greater haplotype diversity, and higher mean sequence divergence in both ITS1 and ITS2. The explanation of this geographic pattern should be addressed taking into consideration more localities from both native and introduced ranges and other data regarding the ecology and historical processes of *E. directus* (J. Vierna et al., in preparation).

The detection of intragenomic divergence in both ITS in a majority of individuals let us conclude that ITS intragenomic divergence is widespread in *E. directus* populations. Interestingly, intragenomic mean sequence divergences in *E. directus* were in some cases higher than the mean sequence divergence for Danish or Canadian ITS sequences. Furthermore, the maximum *p*-distances obtained for ITS1 (0.052 ± 0.010) and ITS2 (0.034 ± 0.011) were considerably higher than the mean sequence divergence between Long Pond and the three Danish localities (Table 3). Similarly, Fairley et al. (2005) found that ITS1 sequence divergences within individuals were in some cases equal to or greater than those within localities in a study of the insect *Anopheles aquasalis*. In any nrDNA region under study, mean intragenomic divergence should be equivalent to mean divergence within populations (or localities forming a population). Nevertheless, estimates of mean intragenomic diver-

Fig. 3. Maximum likelihood (ML) majority-rule consensus trees of *Ensis siliqua*, *E. macha*, *E. directus*, and *E. magnus* internal transcribed spacers (ITS1 and ITS2) reconstructed following the GTR+G model. Bootstrap values ≥ 50% obtained from the ML analysis are indicated above the nodes. When the same subtree was recovered in the maximum parsimony (MP) analysis, the corresponding bootstrap value (≥50%) is indicated below the node. *Ensis directus* sampling localities are indicated in parentheses. Gaps were treated as unknown characters in ML analyses and as fifth character states in MP analyses. (*a*) ITS1 phylogenetic tree. (*b*) ITS2 phylogenetic tree.

Table 3. Mean sequence divergence (*p*-distance, mean ± SE) between species and between localities for both internal transcribed spacers (ITS1 and ITS2).

		ITS1		ITS2	
		Gaps considered	Gaps excluded	Gaps considered	Gaps excluded
sp	*E. directus* – *E. macha**	0.108±0.013	0.093±0.013	0.108±0.018	0.109±0.018
sp	*E. directus* – *E. siliqua**	0.208±0.018	0.195±0.019	0.181±0.023	0.179±0.023
sp	*E. directus* – *E. magnus**	0.218±0.019	0.203±0.019	0.170±0.022	0.170±0.022
sp	*E. macha* – *E. siliqua**	0.193±0.018	0.173±0.018	0.171±0.022	0.168±0.023
sp	*E. macha* – *E. magnus**	0.203±0.019	0.180±0.018	0.173±0.022	0.172±0.023
sp	*E. magnus* – *E. siliqua**	0.017±0.006	0.019±0.006	0.047±0.012	0.045±0.011
sp	*E. directus* – *E. macha*†	0.120±0.012	0.108±0.013	0.120±0.018	0.116±0.018
loc	Long Pond – Sillerslev	0.031±0.007	0.028±0.006	0.021±0.006	0.021±0.006
loc	Long Pond – Sundsøre	0.029±0.007	0.027±0.006	0.013±0.005	0.013±0.005
loc	Long Pond – Juvre Deep	0.027±0.007	0.025±0.006	0.019±0.006	0.019±0.006
loc	Sillerslev – Sundsøre	0.017±0.003	0.015±0.004	0.017±0.005	0.018±0.005
loc	Sillerslev – Juvre Deep	0.015±0.003	0.013±0.003	0.016±0.005	0.015±0.005
loc	Sundsøre – Juvre Deep	0.014±0.003	0.014±0.003	0.018±0.005	0.017±0.005

Note: Mean sequence divergence was calculated with gaps considered and with gaps excluded; the latter values are cited in the text.

*Alignment was performed considering all species and was corrected in Gblocks. ITS1 alignment was reduced to 76% of its original length. ITS2 alignment was reduced to 81% of its original length.

†Alignment was performed considering only *E. directus* and *E. macha* sequences, with no correction in Gblocks (see main text).

gence (the values we obtain, as we cannot sample the entire population) may be much higher than "real" mean divergence if only a reduced number of clones are sampled per individual. This could have happened with individuals Sundsøre 67 and Juvre Deep 119. Thus, though intragenomic divergence occurs in *E. directus* individuals, it could have been overestimated in some of them.

However, the existence of some intragenomic divergence in ITS does not necessarily mean that these spacers are uninformative markers in molecular population studies. For example, Rodriguez-Lanetty and Hoegh-Guldberg (2002) compared the levels of intragenomic divergence in both ITS in the scleractinian coral *Plesiastrea versipora* with the divergence among populations. They found low levels of intragenomic divergence that were always lower than levels of divergence among populations. Therefore, they concluded that ITS were suitable molecular markers for that phylogeographic survey. In the insect *Pediculus humanus*, ITS2 displayed very high intragenomic divergence, so Leo and Barker (2002) concluded that it was an unsuitable marker for molecular population studies in that species.

Our results clearly show that both ITS were informative molecular markers at the population level regardless of the intragenomic divergence detected. Remarkably, although mean intragenomic divergences or maximum *p*-distances were sometimes higher than divergences between localities, ITS1 and ITS2 were able to differentiate each geographic area in both the MP cladogram — based on indel polymorphism — and the ML tree (Figs. 1–2). According to ITS1 and ITS2, both geographic areas are different populations, and this means that individuals from Denmark may have come from a source population other than Long Pond.

However, neither ITS nucleotide polymorphism nor indel polymorphism were informative in differentiating individuals belonging to each Danish locality. The most plausible explanation for this is that individuals from the three Danish localities form a single population, but this should be confirmed by other molecular markers. *Ensis* razor shells have external fertilization and undergo indirect development, so it is not unexpected that sampling localities that are separated by many kilometres are actually the same population.

ITS as molecular markers at the species level

In *Ensis* phylogenetic analyses, species node support was higher under MP than under ML criteria, with the exception of *E. magnus* (ITS1 tree). Similarly to studies done in the Unionidae bivalves (Källersjö et al. 2005), we found that the use of gaps as character states increased the node support for MP phylogenies in the majority of cases (the exception is *E. magnus* in the ITS1 analysis). All ITS1 and ITS2 sequences belonging to each species were monophyletic (except *E. siliqua* sequences in the ITS1 ML tree), and both ITS recovered American and European species as reciprocally monophyletic. This is in accordance with morphological characters (Cosel 2009) and with the mean sequence divergence obtained between species (Table 2). Our results support the conclusion that both ITS are informative phylogenetic markers within genus *Ensis*, but the real ability of these spacers to resolve the phylogeny should be confirmed by other nuclear and mitochondrial sequences, and including all extant species. It should be taken into account that it is

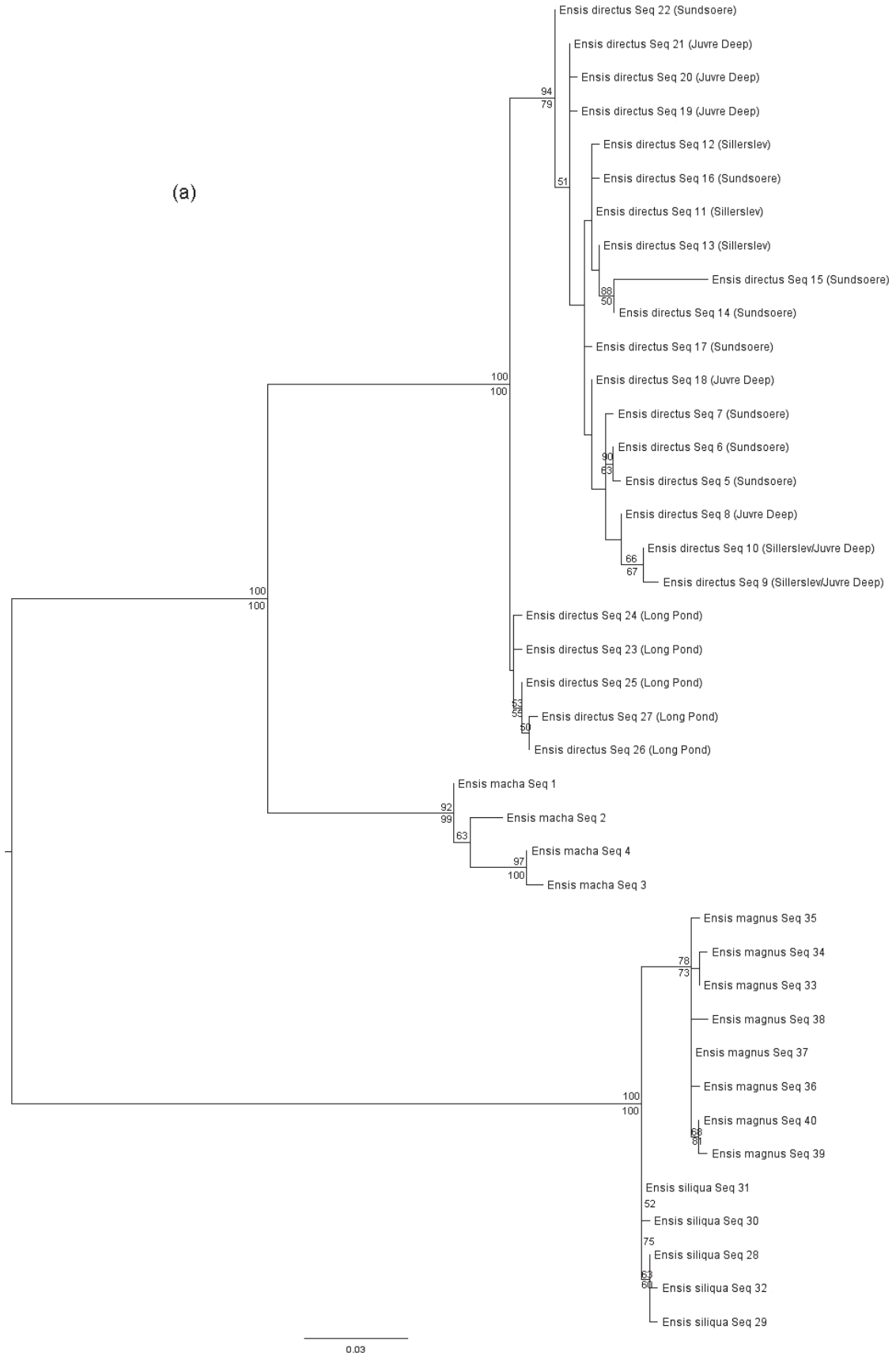
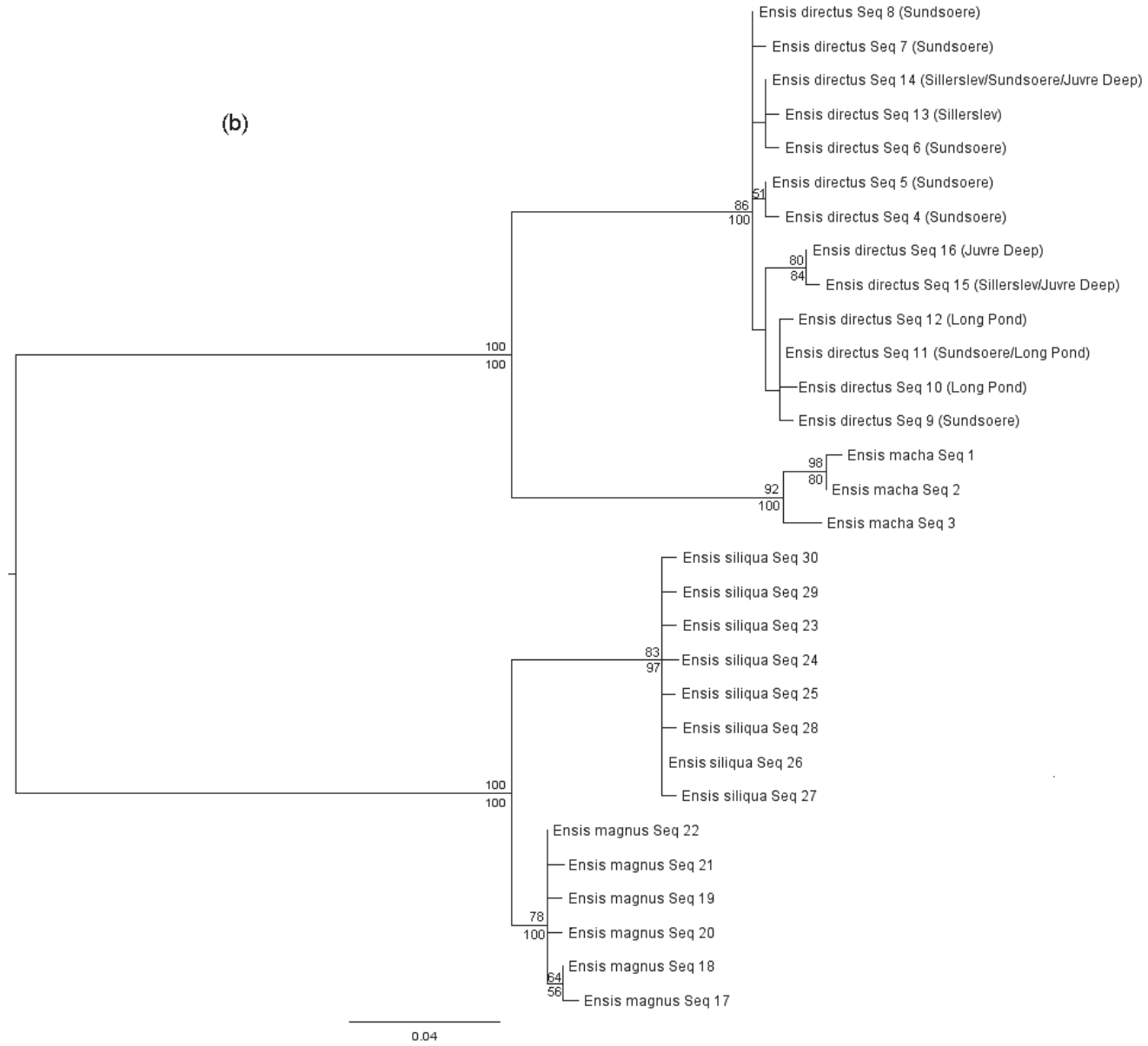

(a)
Ensis directus Seq 22 (Sundsoere)
Ensis directus Seq 21 (Juvre Deep)
Ensis directus Seq 20 (Juvre Deep)
Ensis directus Seq 19 (Juvre Deep)
Ensis directus Seq 12 (Sillerslev)
Ensis directus Seq 16 (Sundsoere)
Ensis directus Seq 11 (Sillerslev)
Ensis directus Seq 13 (Sillerslev)
Ensis directus Seq 15 (Sundsoere)
Ensis directus Seq 14 (Sundsoere)
Ensis directus Seq 17 (Sundsoere)
Ensis directus Seq 18 (Juvre Deep)
Ensis directus Seq 7 (Sundsoere)
Ensis directus Seq 6 (Sundsoere)
Ensis directus Seq 5 (Sundsoere)
Ensis directus Seq 8 (Juvre Deep)
Ensis directus Seq 10 (Sillerslev/Juvre Deep)
Ensis directus Seq 9 (Sillerslev/Juvre Deep)
Ensis directus Seq 24 (Long Pond)
Ensis directus Seq 23 (Long Pond)
Ensis directus Seq 25 (Long Pond)
Ensis directus Seq 27 (Long Pond)
Ensis directus Seq 26 (Long Pond)
Ensis macha Seq 1
Ensis macha Seq 2
Ensis macha Seq 4
Ensis macha Seq 3
Ensis magnus Seq 35
Ensis magnus Seq 34
Ensis magnus Seq 33
Ensis magnus Seq 38
Ensis magnus Seq 37
Ensis magnus Seq 36
Ensis magnus Seq 40
Ensis magnus Seq 39
Ensis siliqua Seq 31
Ensis siliqua Seq 30
Ensis siliqua Seq 28
Ensis siliqua Seq 32
Ensis siliqua Seq 29
0.03

Fig. 3. (*concluded*).



possible to get a species-specific clustering pattern if multiple divergences have occurred since the species last shared a common ancestor. That means that the pattern obtained could be due to the fact that the species may not be very closely related. If this were true, ITS sequences could show a paraphyletic pattern when more closely related species are considered. Though to date we cannot be sure whether *E. siliqua–E. magnus* and *E. directus–E. macha* are in fact sister species, we think they may be quite closely related, as they share some morphological features that distinguish them from other *Ensis* species. However, this point will not be clarified until more data on the phylogeny of genus *Ensis* become available.

Long-term evolution of *Ensis* ITS1–5.8S–ITS2

Considering the levels of intragenomic divergence detected in *E. directus* individuals, it seems that concerted evolution is not completely efficient. We sampled some ITS1 and ITS2 clones from individuals Sundsøre 67, Juvre Deep 103, Sillerslev 8, and Juvre Deep 119 that displayed some point mutations and indels that differentiated them from the other *E. directus* sequences. Though we cannot be completely sure, these clones do not appear to be pseudogenes because their corresponding 5.8S did not display any point mutation or indel. They may be copies that are not yet homogenized but that will eventually undergo homogenization through the mechanisms leading to concerted evolution. Considering all four *Ensis* species, tree topologies support a concerted evolution scenario, as ITS variants from each *Ensis* species are monophyletic. However, the long-term evolution of *Ensis* ITS1–5.8S–ITS2 can also be explained by other evolutionary models. Under a birth-and-

death model, genes are shared for prolonged periods between species, except in exceptional cases when rapid turnover occurs. In these cases, rapid gene turnover can lead to the creation of species-specific gene clusters as a result of frequent gene duplication and loss. Consequently, few or no genes are shared between species (Rooney 2004). In the light of our results, birth-and-death processes and purifying selection may explain *Ensis* ITS1–5.8S–ITS2 long-term evolution. The monophyly of ITS sequences within each species may be the result of these rapid gene duplications and losses, whereas the evolution of the 5.8S may be driven by purifying selection. Sequence clusters observed within both ITS in *E. directus* (Fig. 2) may represent new copies arising by gene duplication along the evolutionary history of this species. In addition, the clear separation of ITS sequences belonging to each geographic area may be a consequence of this rapid gene turnover. Finally, a third possibility that may explain the long-term evolution of *Ensis* ITS1–5.8S–ITS2 is a mixed process of concerted and birth-and-death evolution, as described by Nei and Rooney (2005). According to this model, homogenizing mechanisms (unequal crossovers and gene conversions) and birth-and-death processes (gene duplication and loss) drive the long-term evolution of a given multigene family. Even though our results do not preferentially support any of the three models discussed, we believe that the most logical explanation is that homogenizing mechanisms, birth-and-death processes, and selection are all acting and shaping *Ensis* ITS1–5.8S–ITS2 extant variation.

The nontranscribed spacer region (NTS) of *Ensis* 5S ribosomal DNA was found to evolve under a birth-and-death model, but it was suggested that homogenizing mechanisms may be also taking part within each 5S ribosomal DNA variant in each species (Vierna et al. 2009). The organisation of 5S ribosomal DNA is more flexible than the organisation of the major ribosomal genes (and spacers), as it can be dispersed throughout the genome, found in its own tandem arrays, or found in both types of arrangement (Rooney and Ward 2005). Taking all this together, the long-term evolution of these two ribosomal families in *Ensis* could be reconciled under a mixed process of concerted evolution, birth-and-death evolution, and purifying selection (in the case of the 5S and 5.8S genes). Under this model, homogenizing mechanisms are more efficient within ITS1–5.8S–ITS2, as the major ribosomal genes may not be as dispersed as the 5S ribosomal DNA in *Ensis* genomes, and they may be organised in a smaller number of arrays. Eickbush and Eickbush (2007) pointed out that unequal crossovers are more frequent between sister chromatids than between chromosomes (homologous and nonhomologous). This could be the reason why intragenomic divergence is much higher in *Ensis* NTS than in ITS. Nevertheless, more analyses should be performed to understand the chromosomal organisation of these ribosomal families in *Ensis* razor shells.

Conclusions

From this work, we conclude that (1) ITS1 and ITS2 are suitable molecular markers in *E. directus* despite the intragenomic divergence detected. Both indel and nucleotide polymorphisms are informative at the population level. (2) ITS1 and ITS2 may also be informative markers at the species level. Gaps are useful in reconstructing the phylogenetic relationships among *Ensis* species under MP. (3) The long-term evolution of *Ensis* ITS1–5.8S–ITS2 can be reconciled with the long-term evolution of *Ensis* 5S ribosomal DNA under a mixed process of concerted evolution, birth-and-death evolution, and selection in which the homogenizing mechanisms are less efficient within the 5S ribosomal DNA.

Finally, taking into consideration this and other studies, we recommend analysing the levels of intragenomic divergence before using any nrDNA region as a molecular marker.

Acknowledgements

We are grateful to K. Thomas Jensen, Anne S. Lousdal, and Ray J. Thompson for providing us with the *E. directus* samples, and to Marta Duyos Míguez for reviewing the English grammar. We give sincere thanks to Rudo von Cosel for identifying some of the specimens used in this work, and his comments on *Ensis* taxonomy. Finally, we would like to thank two anonymous reviewers that greatly improved the quality of this article with their comments. J.V. is supported by a "María Barbeito" fellowship from Xunta de Galicia (Spain).

References

Alvarez, B., Krishnan, M., and Gibb, K. 2007. Analysis of intragenomic variation of the rDNA internal transcribed spacers (ITS) in *Halichondrida* (Porifera: Demospongiae). J. Mar. Biol. Assoc. U.K. **87**(6): 1599–1605. doi:10.1017/S0025315407058407.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**(4): 540–552. PMID:10742046.

Cheng, H.L., Xia, D.Q., Wu, T.T., Meng, X.P., Ji, H.J., and Dong, Z.G. 2006. Study on sequences of ribosomal DNA internal transcribed spacers of clams belonging to the Veneridae family (Mollusca: Bivalvia). Acta Genet. Sin. **33**(8): 702–710. doi:10.1016/S0379-4172(06)60102-9. PMID:16939004.

Chow, S., Ueno, Y., Toyokawa, M., Oohara, I., and Takeyama, H. 2009. Preliminary analysis of length and GC content variation in the ribosomal first internal transcribed spacer (ITS1) of marine animals. Mar. Biotechnol. (NY), **11**(3): 301–306. doi:10.1007/s10126-008-9153-2. PMID:18937008.

Cosel, R. von 2009. The razor shells of the eastern Atlantic, part 2. Pharidae II: the genus *Ensis* Schumacher, 1817 (Bivalvia, Solenoidea). Basteria, **73**: 9–56.

Cosel, R. von, Dörjes, J., and Mühlenhardt-Siegel, U. 1982. Die amerikanische schwertmuschel *Ensis directus* (Conrad) in der Deutschen Bucht. I. Zoogeographie und taxonomie mi vergleich mit den einheimischen schwertmuschel-Arten. Senckenb. Marit. **14**: 147–173.

Eickbush, T.H., and Eickbush, D.G. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics, **175**(2): 477–485. doi:10.1534/genetics.107.071399. PMID: 17322354.

Fairley, T.L., Kilpatrick, C.W., and Conn, J.E. 2005. Intragenomic heterogeneity of internal transcribed spacer rDNA in neotropical malaria vector *Anopheles aquasalis* (Diptera: Culicidae). J. Med. Entomol. **42**(5): 795–800. doi:10.1603/0022-2585(2005)042[0795:IHOITS]2.0.CO;2. PMID:16365998.

Felsenstein, J. 1985. Confidence limits on phylogenies: an ap-

proach using the bootstrap. Evolution, **39**(4): 783–791. doi:10.2307/2408678.

Fernández, A., García, T., Asensio, I., Rodríguez, M.A., González, I., Hernández, P.E., and Martín, R. 2001. PCR-RFLP analysis of the internal transcribed spacer (ITS) region for identification of 3 clam species. J. Food Sci. **66**(5): 657–661. doi:10.1111/j.1365-2621.2001.tb04617.x.

Fujiwara, M., Inafuku, J., Takeda, A., Watanabe, A., Fujiwara, A., Kohno, S., and Kubota, S. 2009. Molecular organization of 5S rDNA in bitterlings (Cyprinidae). Genetica, **135**(3): 355–365. doi:10.1007/s10709-008-9294-2. PMID:18648989.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**(7): 685–695. PMID:9254330.

Guindon, S., and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**(5): 696–704. doi:10.1080/10635150390235520. PMID:14530136.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41**: 95–98.

Harris, D.J., and Crandall, K.A. 2000. Intragenomic variation within ITS1 and ITS2 of freshwater crayfishes (Decapoda: Cambaridae): implications for phylogenetic and microsatellite studies. Mol. Biol. Evol. **17**(2): 284–291. PMID:10677851.

Heath, D.D., Rawson, P.D., and Hilbish, T.J. 1995. PCR-based nuclear markers identify alien blue mussel (*Mytilus* spp.) genotypes on the west coast of Canada. Can. J. Fish. Aquat. Sci. **52**(12): 2621–2627. doi:10.1139/f95-851.

Insua, A., López-Piñón, M.J., Freire, R., and Méndez, J. 2003. Sequence analysis of the ribosomal DNA internal transcribed spacer region in some scallop species (Mollusca: Bivalvia: Pectinidae). Genome, **46**(4): 595–604. doi:10.1139/g03-045. PMID: 12897868.

Jarvis, J.P., Luedeman, J.K., and Shier, D.R. 1983. Comments on computing the similarity of binary trees. J. Theor. Biol. **100**: 427–433.

Källersjö, M., von Proschwitz, T., Lundberg, S., Eldenäs, P., and Erséus, C. 2005. Evaluation of *ITS* rDNA as a complement to mitochondrial gene sequences for phylogenetic studies in freshwater mussels: an example using Unionidae from north-western Europe. Zool. Scr. **34**(4): 415–424. doi:10.1111/j.1463-6409.2005.00202.x.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics, **23**(21): 2947–2948. doi:10.1093/bioinformatics/btm404. PMID:17846036.

Leo, N.P., and Barker, S.C. 2002. Intragenomic variation in ITS2 rDNA in the louse of humans, *Pediculus humanus*: ITS2 is not a suitable marker for population studies in this species. Insect Mol. Biol. **11**(6): 651–657. doi:10.1046/j.1365-2583.2002.00367.x. PMID:12421423.

Mahidol, C., Na-Nakorn, U., Sukmanomon, S., Yoosuk, W., Taniguchi, N., and Nguyen, T.T.T. 2007. Phylogenetic relationships among nine scallop species (Bivalvia: Pectinidae) inferred from nucleotide sequences of one mitochondrial and three nuclear gene regions. J. Shellfish Res. **26**(1): 25–32. doi:10.2983/0730-8000(2007)26[25:PRANSS]2.0.CO;2.

Nei, M., and Rooney, A.P. 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. **39**(1): 121–152. doi:10.1146/annurev.genet.39.073003.112240. PMID: 16285855.

Posada, D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. **25**(7): 1253–1256. doi:10.1093/molbev/msn083. PMID:18397919.

Rodriguez-Lanetty, M., and Hoegh-Guldberg, O. 2002. The phylogeography and connectivity of the latitudinally widespread scleractinian coral *Plesiastrea versipora* in the Western Pacific. Mol. Ecol. **11**(7): 1177–1189. doi:10.1046/j.1365-294X.2002.01511.x. PMID:12074725.

Rooney, A.P. 2004. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans. Mol. Biol. Evol. **21**(9): 1704–1711. doi:10.1093/molbev/msh178. PMID:15175411.

Rooney, A.P., and Ward, T.J. 2005. Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proc. Natl. Acad. Sci. U.S.A. **102**(14): 5084–5089. doi:10.1073/pnas.0409689102. PMID: 15784739.

Rozas, J., Sánchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics, **19**(18): 2496–2497. doi:10.1093/bioinformatics/btg359. PMID:14668244.

Swofford, D.L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Sword, G.A., Senior, L.B., Gaskin, J.F., and Joern, A. 2007. Double trouble for grasshopper molecular systematics: intra-individual heterogeneity of both mitochondrial 12S-valine-16S and nuclear internal transcribed spacer ribosomal DNA sequences in *Hesperotettix viridis* (Orthoptera: Acrididae). Syst. Entomol. **32**(3): 420–428. doi:10.1111/j.1365-3113.2007.00385.x.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24**(8): 1596–1599. doi:10.1093/molbev/msm092. PMID:17488738.

Toro, J., Innes, D.J., and Thompson, R.J. 2004. Genetic variation among life-history stages of mussels in a *Mytilus edulis–M. trossulus* hybrid zone. Mar. Biol. (Berl.), **145**: 713–725. doi:10.1007/s00227-004-1363-1.

Vierna, J., González-Tizón, A.M., and Martínez-Lage, A. 2009. Long-term evolution of 5S ribosomal DNA seems to be driven by birth-and-death processes and selection in *Ensis* razor shells (Mollusca: Bivalvia). Biochem. Genet. **47**(9–10): 635–644. doi:10.1007/s10528-009-9255-1. PMID:19633948.

Wörheide, G., Hooper, J.N.A., and Degnan, B.M. 2002. Phylogeography of western Pacific *Leucetta 'chagosensis'* (Porifera: Calcarea) from ribosomal DNA sequences: implications for population history and conservation of the Great Barrier Reef World Heritage Area (Australia). Mol. Ecol. **11**(9): 1753–1768. doi:10.1046/j.1365-294X.2002.01570.x. PMID:12207725.

Wörheide, G., Nichols, S.A., and Goldberg, J. 2004. Intragenomic variation of the rDNA internal transcribed spacers in sponges (phylum Porifera): implications for phylogenetic studies. Mol. Phylogenet. Evol. **33**(3): 816–830. doi:10.1016/j.ympev.2004.07.005. PMID:15522806.

Yu, E.T., Juinio-Meñez, M.A., and Monje, V.D. 2000. Sequence variation in the ribosomal DNA internal transcribed spacer of *Tridacna crocea.* Mar. Biotechnol. (NY), **2**(6): 511–516. doi:10.1007/s101260000033. PMID:14961174.

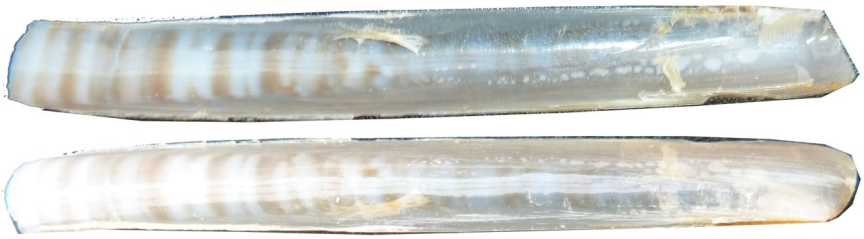
Zhang, J., Dyer, K.D., and Rosenberg, H.F. 2000. Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. Proc. Natl. Acad. Sci. U.S.A. **97**(9): 4701–4706. doi:10.1073/pnas.080071397. PMID: 10758160.

10.3 Photographs of the valves of the currently known Atlantic *Ensis* species

Ensis ensis (Linné, 1758)

bar = 2 cm

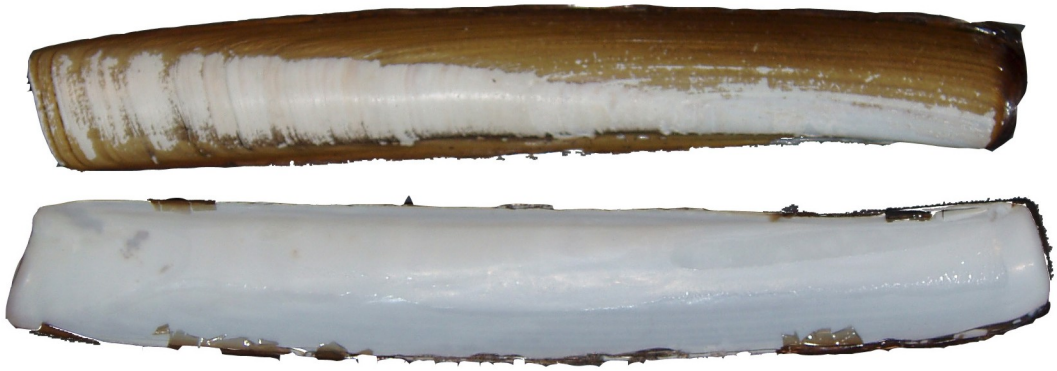
Ensis minor (Chenu, 1843)



Ensis magnus Schumacher, 1817

bar = 2 cm

Ensis siliqua (Linné, 1758)



bar = 2 cm

Ensis goreensis (Clessin, 1888)

bar = 2 cm

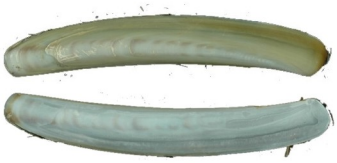
Ensis directus (Conrad, 1843)

bar = 2 cm

Ensis megistus megistus Pilsbry and McGinty, 1943

bar = 2 cm

Ensis megistus coseli Vierna, 2013



bar = 2 cm

Ensis terranovensis Vierna and Martínez-Lage, 2012

bar = 2 cm

Ensis macha (Molina, 1792)

bar = 2 cm

This thesis is dedicated to my family.

A mi padre, a mi madre y a mi abuela Saruca.

En recuerdo del abuelo Rabudo, de la abuela Fefa y del abuelo Josecho.